NOVEL TRANSFORMER ARCHITECTURES FOR 3D MULTI-MODAL AND MULTI-ORGAN MEDICAL IMAGE SEGMENTATION

by

CHENGYIN LI

DISSERTATION

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2024

MAJOR: COMPUTER SCIENCE

Approved By:

Advisor Date

Committee Member 1 Date

Committee Member 2 Date

Committee Member 3 Date

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my PhD advisor, Dr. Dongxiao Zhu, whose support, insightful discussions, and guidance have been instrumental in the completion of this dissertation. His patience, knowledge, and encouragement have been invaluable.

I also extend my heartfelt thanks to my PhD committee members, Dr. Nathan Fisher, Dr. Prashant Khanduri, and Dr. Jitao Zhang, for their career advice and constructive feedback.

Special thanks go to Dr. Indrin Julian Chetty, Dr. Kundan S. Thind, and Dr. Hassan Bagher-Ebadian for their assistance in providing computational resources, medical image data, and for their contributions to our research discussions.

I am also grateful to Yao Qiang, Rafi Ibn Sultan, Hui Zhu, Xiangyu Zhou, Dr. Xin Li, Dr. Deng Pan, Dr. Xiangrui Li, and all members of the Trustworthy AI Lab for their companionship, support, and wisdom throughout this journey.

This dissertation was supported by NIH grants R33HD105610 and R61HD105610, the Thomas Rumble Fellowship provided by Wayne State University, and funding from Henry Ford Health System. I am grateful for their support and contributions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

This chapter provides an introduction to medical image segmentation-related background, challenges, and deep learning (DL) techniques. It also includes an overview of the main results of this dissertation.

1.1 Background

Medical image analysis, a critical research area within medical imaging, leverages computer vision techniques to generate complex visual representations of the human body, crucial for disease diagnosis and therapy [2, 38, 159, 72, 96]. This field has been transformed by the integration of various image acquisition tools, such as Magnetic Resonance (MR) Imaging [4, 84], Computed Tomography (CT) [50, 71], ultrasound [132], and X-ray imaging [97], which are essential in non-invasive and minimally invasive diagnostic methods. The core of medical image analysis involves extracting semantic information from these images, facilitating both abstract interpretation and quantitative measurements [38, 96]. Key processes in this field [38, 159] include feature extraction (identifying essential characteristics from data), segmentation (dividing images into specific regions for detailed analysis), classification (grouping data based on shared attributes), registration (aligning data from multiple sources), and measurement (obtaining precise quantitative values for diagnosis and treatment evaluation). These advancements have significantly impacted diagnostic radiology and minimally invasive therapy, underscoring the importance of imaging modality choice in clinical diagnosis and understanding disease progression.

This dissertation mainly examines CT [12], also known as Computed Axial Tomography, which is a cornerstone in medical imaging, emphasizing its critical role in providing detailed insights into the body's internal structures for accurate anatomical analysis and diagnosis. CT technology, which has advanced significantly since its introduction in the early 1970s, generates cross-sectional images through computer-processed X-ray measurements from multiple angles, enabling a non-invasive internal view [114]. These images are distinguished by their high contrast and high spatial resolution, effectively differentiating between tissues of similar density and offering clear visualization of small structures. The scanning process, quick and minimally discomforting, involves the patient being exposed to X-ray beams with the attenuated rays captured and converted into detailed cross-sectional and 3D images. The rapid imaging capability of CT is especially crucial in emergencies. Its high-resolution outputs are instrumental in distinguishing normal anatomy from pathological conditions, effectively identifying tumors [52], bone abnormalities [8], and lung issues [144], as well as aiding in the guidance and monitoring of treatments such as biopsies, chemotherapy, and radiation therapy [112, 132].

1.2 Medical Image Segmentation

This dissertation focuses on medical image segmentation. Image segmentation in medical diagnostics and treatment represents a fundamental process in medical imaging, where a medical image is partitioned into multiple segments or sets of pixels (voxels), transforming its representation into a format that is both simpler and more meaningful for analysis [90, 3, 72]. Segmentation enhances diagnostic accuracy by enabling the precise delineation of anatomical structures and pathological regions. Such precision is indispensable for accurately identifying various diseases, including tumors, vascular diseases, and musculoskeletal disorders [68]. In the realm of treatment planning [117], accurate segmentation proves vital, particularly in disciplines like radiotherapy, where it is essential to precisely outline target treatment areas while sparing critical structures. Lastly, precise medical image segmentation, essential in personalized medicine, facilitates customized treatments tailored to patients' unique anatomical and pathological profiles, enabling more effective, patient-specific interventions [106].

1.2.1 Evolution from Manual to Automated Techniques

The evolution of medical image segmentation began with manual techniques, where radiologists delineated images by hand, which faced time and accuracy limitations [76]. To mitigate these issues, semi-automatic methods like thresholding and region growing were introduced, blending manual input with automated algorithms for enhanced speed and accuracy, yet still necessitating human involvement [108, 100]. The advent of fully automated methods marked a significant advancement, propelled by strides in computer technology and software development. These methods, including edge detection and clustering, reduced human input and enabled complex tasks like 3D volume rendering, driven by advancements in imaging modalities [108]. A pivotal moment in this evolution was the integration of machine learning, particularly deep learning, into medical image segmentation.

1.2.2 Deep Learning Techniques

Convolutional Network-Based Approaches Convolutional Neural Networks (CNNs) have been instrumental in the advancement of medical image segmentation [60]. The foundational element of CNNs is the convolutional layer, which applies filters to input images to extract features [65]. As data progresses through the network, these features become increasingly abstract, encapsulating higher-level concepts. In segmentation tasks, CNNs initially identify edges, textures, and other salient features in the early layers. Deeper layers, in contrast, are tasked with recognizing more intricate structures that are critical for specific segmentation tasks [60, 1].

A significant milestone in employing CNNs for medical image segmentation is the introduction of the U-Net architecture [110]. U-Net, specifically designed for biomedical image segmentation,

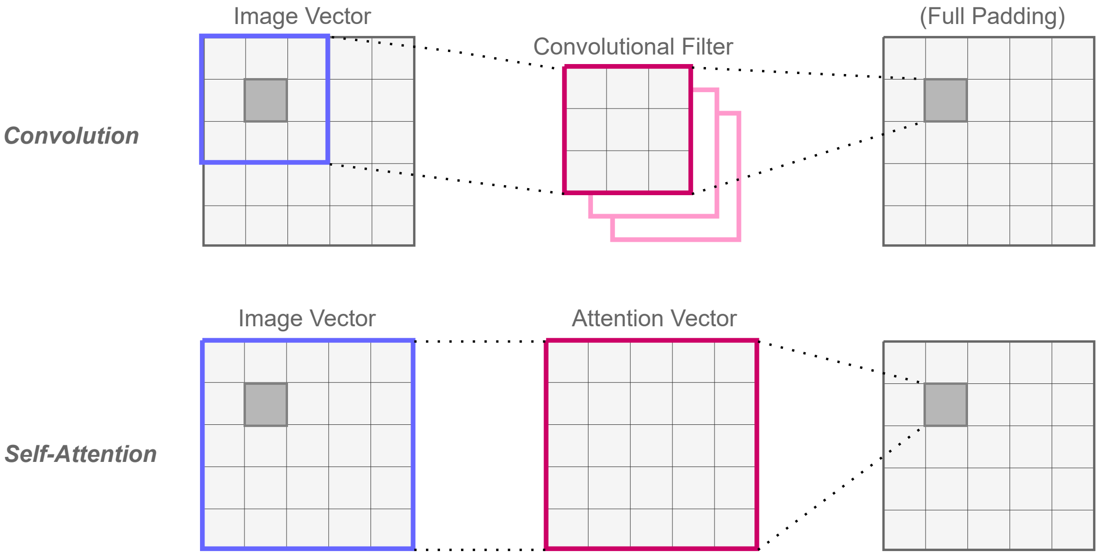

Figure 1: Convolution operations and self-attention mechanisms access regions of vastly different sizes. In the context of vision, self-attention is specifically designed to learn the relationships between one pixel and all other positions, including those far apart, enabling it to capture global dependencies effectively. Conversely, convolution is governed by the size of the convolutional filters and primarily focuses on local regions.

features a symmetrical structure with contracting and expanding paths, facilitating precise localization. This architecture is particularly effective in tasks requiring both contextual understanding and precise localization, such as tumor segmentation or organ delineation in medical imagery. Over time, several adaptations and enhancements of the original U-Net model have emerged [164, 101]. Notable developments include 3D U-Nets [23, 95], optimized for volumetric data, and attention U-Nets [101], which incorporate attention mechanisms to concentrate on specific regions of interest. Additionally, nnUNet [56] has been developed to address variability, automating dataset-specific processing and optimizing the training process.

While CNNs based approaches have achieved impressive results in medical image segmentation, they tend to focus on local features due to their convolutional nature (as shown in Figure 1), which might lead to suboptimal performance in cases where global context or long-range dependencies are crucial [20, 123].

Transformer Based Approaches The introduction of Transformer-based models in medical image segmentation marks a significant shift away from traditional convolutional methods [20, 48, 14, 120]. Initially developed for natural language processing tasks, Transformers have been effectively adapted to navigate the complexities of medical imaging data, thereby offering new perspectives and enhanced capabilities in segmentation tasks [72]. At the core of Transformers lies their reliance on self-attention mechanisms for data processing [128], a feature that is particularly advantageous in the context of medical imaging (as shown in Figure 1). This ability to capture long-range dependencies and contextual information across entire images is crucial for understanding the complex spatial relationships and diverse anatomical structures found in medical images [72].

TransUNet [20] stands as one of the pioneers in successfully employing the Vision Transformer (ViT) for medical image segmentation, utilizing pre-trained weights from image classification. In this architecture, convolutional layers form the core for feature extraction, while transformers are employed to grasp the long-range global context. Following this approach, several studies [123, 148, 18] have emerged; however, they often find that multiple transformer layers alone are insufficient to capture long-term dependencies alongside precise spatial information within hierarchical feature maps.

To tackle this limitation, researchers have introduced self-attention mechanisms into the convolution operation [137, 157, 42] to enhance medical image segmentation. For instance, Gao et al. [42] integrated self-attention into a CNN to improve segmentation outcomes. Zhou et al. [157] proposed a hybrid model that interweaves convolution and self-attention in both the encoder and decoder modules. While these methods demonstrate improved performance, their intricate design of convolution and self-attention modules can limit the scalability of developing more advanced transformer architectures.

More recently, the Swin Transformer [85] has shown promising results, demonstrating linear complexity in self-attention calculations. This approach facilitates efficient learning of long-range contexts and the generation of hierarchical feature maps. Building on this, SwinUNet [13] employs hierarchical transformer blocks within a U-Net-like architecture. DS-TransUNet [78] introduces a more parallel encoder to process inputs at different resolutions, and SwinUNETR [120] leverages pre-training on a vast medical image dataset.

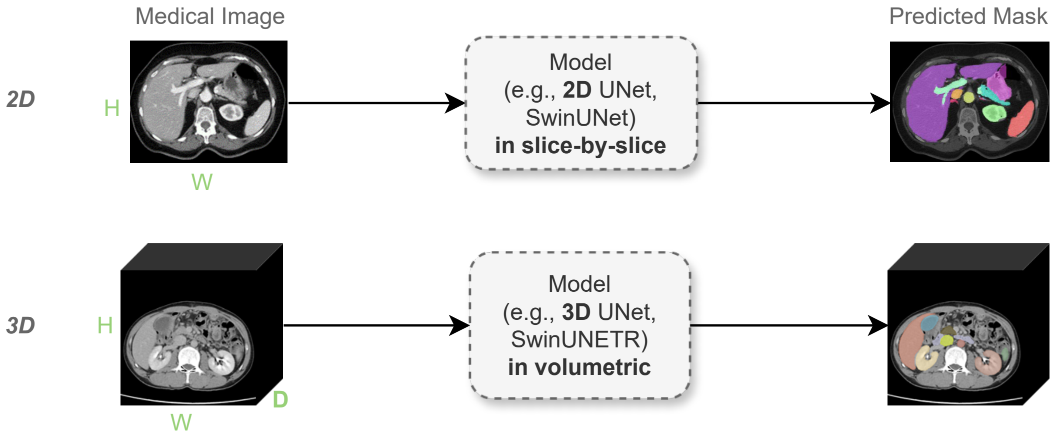

Figure 2: Comparison of 2D (slice-by-slice) and 3D (volumetric) Architectures for Medical Image Segmentation.

2D and 3D Medical Image Segmentation 2D and 3D medical image segmentation are critical techniques in medical imaging that facilitate precise analysis and diagnosis [111, 23, 95, 101, 20, 48, 120]. Figure 2 shows a comparison of 2D and 3D architectures for medical image segmentation. The top row illustrates a 2D model segmenting a medical image slice-by-slice, resulting in a predicted mask for each individual slice. The bottom row shows a 3D model performing volumetric segmentation on the entire medical image volume, producing a comprehensive predicted mask across the entire volume input. The dimensions H, W, and D represent height, width, and depth, respectively.

2D segmentation focuses on slice-by-slice analysis, making it less computationally intensive and easier to implement for tasks involving single organs or regions with clear boundaries. This method is particularly effective in cases where the target structures are primarily 2D or when high-resolution slices are available. However, 2D segmentation often struggles with capturing the spatial continuity and context of anatomical structures across multiple slices, which can lead to inaccuracies in three-dimensional reconstructions and analyses.

Despite the simplicity of the 2D segmentation setting, we mainly focus on 3D medical image segmentation in this dissertation for the following factors. First, 3D segmentation provides a holistic view of anatomical structures, capturing spatial relationships and continuity across all dimensions. This comprehensive perspective is crucial for accurately assessing complex structures and pathologies that span multiple imaging slices, such as tumors, vascular networks, and organ boundaries. Furthermore, 3D segmentation reduces the risk of slice-to-slice inconsistencies that can occur in 2D segmentation, thereby improving the reliability and reproducibility of the analysis. Advanced 3D techniques also facilitate more precise volumetric measurements and enable sophisticated simulations and visualizations, which are invaluable for planning surgical interventions, radiation therapy, and other treatments. Despite its higher computational demands, the integration of advanced machine learning algorithms and improved hardware capabilities is making 3D segmentation more accessible and efficient. Overall, the enhanced accuracy, continuity, and clinical relevance of 3D segmentation make it a superior choice for many medical imaging applications, driving advancements in patient care and medical research.

4D Medical Image Segmentation Beyond the 2D and 3D paradigms, 4D medical image segmentation incorporates the temporal dimension, allowing for the analysis of dynamic changes within anatomical structures over time. This advanced technique is particularly beneficial for applications such as cardiac imaging, where capturing the heart's motion across different phases of the cardiac cycle is crucial for accurate assessment. By integrating spatial and temporal data, 4D segmentation provides a more comprehensive understanding of organ function and disease progression, enabling precise tracking of tumors or monitoring the efficacy of treatments. Although 4D segmentation presents additional computational challenges and demands sophisticated algorithms for temporal registration and motion correction, its ability to provide detailed insights into dynamic physiological processes makes it an invaluable tool in the advancement of personalized medicine and real-time diagnostic applications.

To effectively handle the additional temporal dimension in 4D medical image segmentation, deep learning methods have been adapted and extended. 3D Convolutional Neural Networks (CNNs) have been enhanced to incorporate temporal information by utilizing 3D+time convolutional layers [126, 154], which allow the network to learn spatiotemporal features directly from the 4D data. Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, are combined with CNNs to capture temporal dependencies, enabling the model to account for changes over time within the anatomical structures [147]. Additionally, U-Net variants have been adapted to handle 4D data by adding temporal convolution layers, maintaining the architecture's efficiency in capturing detailed features while incorporating temporal continuity [34]. These advanced deep learning methods significantly improve the accuracy and robustness of 4D segmentation, making them essential for dynamic and comprehensive medical image analysis.

1.2.3 Training Objectives

Training objective functions are a critical component in developing effective medical image segmentation models, as they guide the optimization process by quantifying the difference between the predicted and ground truth segmentations. Commonly used loss functions include Dice Loss [55, 153] and Cross-Entropy Loss [25]. Dice Loss measures the overlap between predicted and actual segmentation masks, making it particularly useful for handling imbalanced datasets by focusing on regions of interest. Conversely, Cross-Entropy Loss calculates the pixel-wise classification error and is effective for multi-class segmentation tasks. Often, a combination of these losses is employed to leverage their respective strengths, enhancing both global overlap and local accuracy. Below, we provide a detailed overview of several popular loss functions.

CE Loss Cross Entropy (CE) is derived from the Kullback-Leibler (KL) Divergence [25], which quantifies the dissimilarity between two probability distributions, typically denoted as P and Q. The KL Divergence, a statistical distance measure, is defined as:

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) = -\sum_i p_i \log q_i + \sum_i p_i \log p_i = H(P, Q) - H(P), \tag{1.1}$$

where $H(P, Q) = -\sum_i p_i \log q_i$ is the cross entropy between the distribution P and Q, and $H(P) = -\sum_i p_i \log p_i$ is the entropy of distribution P. In typical machine learning scenarios, the data distribution P is assumed to be represented by the training dataset. Minimizing the KL Divergence between the ground truth distribution P and the predicted distribution Q is equivalent to minimizing the Cross-Entropy $H(P, Q)$. The CE loss is defined as follows:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_c^C \sum_i^N g_i^c \log s_i^c, \tag{1.2}$$

where g_i^c is the ground truth binary indicator of class label c of voxel i, and s_i^c is the corresponding predicted segmentation probability. C is the number of classes, and N is the number of pixels/voxels.

Dice Loss Dice loss directly optimizes the Dice Similarity Coefficient (DSC), a widely employed metric for segmentation evaluation. Generally, two variants of the Dice loss are recognized [55]. One variant includes squared terms in the denominator [95], and it is defined as follows:

$$\mathcal{L}_{\text{Dice-square}} = 1 - \frac{2\sum_c^C \sum_i^N s_i^c g_i^c}{\sum_c^C \sum_i^N (s_i^c)^2 + \sum_c^C \sum_i^N (g_i^c)^2}. \tag{1.3}$$

The other does not use the squared terms in the denominator [35], which is defined by

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_c^C \sum_i^N s_i^c g_i^c}{\sum_c^C \sum_i^N s_i^c + \sum_c^C \sum_i^N g_i^c}. \tag{1.4}$$

Unless otherwise specified, we use the no-squared version (Equation 1.4) as the default configuration.

Boundary Loss Boundary loss was introduced to minimize dissimilarities between predicted and ground truth segmentations [88]. The boundary loss function can be expressed as follows:

$$\mathcal{L}_{\text{BD}} = \sum_{\Omega} \phi_G(p) s_\theta(p), \tag{1.5}$$

where $\phi_G(p) = -D_G(q)$ if point $p \in G$, and $\phi_G(p) = D_G(q)$ otherwise. $D_G(q)$ is the distance map of ground truth. An additional distribution-based loss to stabilize the DL model training process often included [64]. Here researchers often use boundary loss in conjunction with the distribution-based CE loss.

$$\mathcal{L}_{\text{CE+BD}} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\mathcal{L}_{\text{BD}}, \tag{1.6}$$

where α is a hyperparameter and can be optimized through empirical studies.

TopK Loss TopK loss is a variation of cross entropy designed to prioritize challenging samples during training. It retains the K percent worst pixels for loss, irrespective of their loss/probability values [135, 88]. It is defined by

$$\mathcal{L}_{\text{TopK}} = -\frac{1}{N}\sum_{c}^{C}\sum_{i\in K} g_i^c \log s_i^c. \tag{1.7}$$

1.2.4 Evaluation Metrics

Evaluating the performance of medical image segmentation models requires a set of robust metrics that can accurately reflect the model's effectiveness in delineating anatomical structures and pathological regions. The following are the key metrics commonly used in this domain.

Dice Similarity Coefficient (DSC) Researchers also refer to the Dice Similarity Coefficient (DSC) as the Dice score. The DSC is a widely used metric that quantifies the overlap between the predicted segmentation mask and the ground truth. It is defined as

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}, \tag{1.8}$$

where A is the predicted mask and B is the ground truth mask. A higher DSC indicates better overlap, with a maximum value of 1 signifying perfect agreement and 0 indicating no overlap. In

2D segmentation, the DSC measures the area overlap between the predicted mask and the ground truth, whereas in 3D segmentation, it measures the overlap through volumes.

Hausdorff Distance (HD) The Hausdorff Distance (HD) is a measure used to determine the similarity between two sets of points. In medical image analysis, it is commonly used to evaluate the accuracy of segmentation algorithms by comparing the segmented region against the ground truth region. The HD is particularly useful because it considers the worst-case scenario by measuring the greatest distance from a point in one set to the closest point in the other set. Given two sets A and B, the Hausdorff Distance $d_H(A, B)$ is defined as:

$$d_H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a)\}, \tag{1.9}$$

where $d(a, b)$ is the Euclidean distance between points a and b. sup and inf denote supremum and infimum, respectively. The 95th percentile of the Hausdorff Distance (HD95) is often used to reduce the impact of outliers.

Average Symmetric Surface Distance (ASSD) The Average Symmetric Surface Distance (ASSD) is another metric used to evaluate the accuracy of segmentation algorithms, particularly in medical imaging. Unlike the Hausdorff Distance, which considers the worst-case scenario, the ASSD provides an average measure of the distance between two surfaces, offering a more balanced assessment of segmentation quality. The ASSD between two sets of points A and B is defined as the average of the minimum distances from each point in one set to the closest point in the other set.

Mathematically, it can be expressed as:

$$\text{ASSD}(A, B) = \frac{1}{|A| + |B|} \left(\sum_{a \in A} \min_{b \in B} d(a, b) + \sum_{b \in B} \min_{a \in A} d(b, a) \right), \tag{1.10}$$

where $d(a, b)$ is the distance between points a and b, $|A|$ and $|B|$ are the number of points in sets A and B respectively. Researchers also use ASD (Average Surface Distance) when considering only the distances from each point in the segmentation surface to the nearest point in the ground truth mask surface, without considering the reverse. In general, for both ASSD and ASD, a smaller value indicates a better match between the segmentation and the ground truth, implying higher segmentation accuracy.

Jaccard Index (Intersection over Union, IoU) The Jaccard Index, also known as Intersection over Union (IoU), is a crucial metric for evaluating the performance of segmentation algorithms, particularly in medical image analysis. It quantifies the similarity between the predicted segmentation and the ground truth by comparing their overlap with their union.

Mathematically, the Jaccard Index for two sets A and B is defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \tag{1.11}$$

where $|A \cap B|$ denotes the number of elements (or pixels) in the intersection of sets A and B, and $|A \cup B|$ denotes the number of elements (or pixels) in the union of sets A and B.

Precision and Recall Precision and recall are crucial metrics used to evaluate the performance of segmentation algorithms, particularly in the context of medical imaging. Precision is defined as the ratio of true positive results to the sum of true positive and false positive results. It measures the accuracy of the positive predictions made by the model. Mathematically, it is expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{1.12}$$

Recall, also known as sensitivity or true positive rate, is the ratio of true positive results to the sum of true positive and false negative results. It measures the model's ability to identify all relevant instances in the dataset. Mathematically, it is expressed as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{1.13}$$

In the context of medical imaging segmentation, precision, and recall provide insights into the algorithm's performance. High precision indicates that the segmentation algorithm produces a low number of false positives, meaning most of the detected regions are correctly identified. High recall indicates that the segmentation algorithm successfully identifies most of the true positive regions, with few false negatives.

F1 Score The F1 Score is a metric used to evaluate the accuracy of a segmentation algorithm by balancing both precision and recall. It is particularly useful when you need a single measure that takes both false positives and false negatives into account, which is crucial in medical imaging where both types of errors can have significant consequences.

The F1 Score is the harmonic mean of precision and recall, and it is calculated as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{1.14}$$

The F1 Score ranges from 0 to 1, where 1 indicates perfect precision and recall (i.e., all predicted positives are true positives, and all actual positives are correctly identified), and 0 indicates the worst performance, where either precision or recall is zero.

In medical imaging segmentation, the F1 Score provides a comprehensive evaluation of the algorithm's performance by considering both over-segmentation (false positives) and under-segmentation (false negatives). A higher F1 Score indicates a better balance between precision and recall, implying higher segmentation accuracy and reliability.

Mean Absolute Distance The Mean Absolute Distance (MAD) is another metric used to evaluate the accuracy of segmentation algorithms, particularly in medical imaging. It measures the average of the absolute differences between corresponding points on the predicted and ground truth surfaces, providing a straightforward assessment of the segmentation quality.

Mathematically, the MAD between two sets of points A and B is defined as:

$$\mathrm{MAD}(A, B) = \frac{1}{|A|} \sum_{a \in A} |d(a)|, \tag{1.15}$$

where $d(a)$ is the distance between a point a in set A and its corresponding point in set B, and $|A|$ is the number of points in set A.

In the context of medical imaging segmentation, the MAD provides a measure of how closely the segmented surface aligns with the ground truth surface. A smaller MAD indicates a better match between the segmentation and the ground truth, implying higher segmentation accuracy. The MAD is particularly useful in applications where it is important to have an intuitive and easily interpretable measure of segmentation accuracy. It complements other metrics like the ASSD or ASD by providing a different perspective on the segmentation quality.

1.2.5 Challenges in Medical Image Segmentation

Limitation of CNN-based Methods in Capturing Long-range Global Context. Medical image segmentation is significantly challenged by the variability inherent in medical images [75]. This variability is due to factors like differences in patient demographics, imaging modalities, and settings. Such diversity complicates the creation of universal segmentation models that perform accurately across varied scenarios. Furthermore, medical images often contain noise and artifacts [61], such as motion artifacts or scanner-induced noise, which considerably affect segmentation accuracy. CNN-based approaches, while effective in medical image segmentation, predominantly focus on local features. This focus can lead to suboptimal performance where the global context or long-range dependencies are essential [20, 123]. CNNs are proficient in local feature extraction but often fall short in understanding the broader image context, which underscores the need for methods that effectively integrate both local and global contextual information [20, 123].

Transformer-based encoders leveraging the self-attention mechanism show great promise in this regard. Transformers were initially designed to capture long-term dependencies of sequential data with stacked self-attention layers [128] and achieved great success in National Language Processing (NLP) tasks. Inspired by this, Dosovitskiy et al. [30] proposed ViT by formulating image classification as a sequence prediction task of the image patch (region) sequence, thereby capturing long-term dependencies within the input image. TransUNet [20] successfully adapts ViT to the medical image segmentation task, where the encoder consists of the Fully Convolutional Network (FCN)-based layers followed by several layers of transformer (multi-head self-attention module) to better capture the global context from medical image inputs. The subsequent studies [123, 148, 18, 58] follow a similar route. However, learning the long-term dependencies that typically contain precise spatial information in low-level feature maps requires more than a few transformer layers for high-level feature inputs. More recently, Swin Transformer [85] demonstrated that it can simultaneously learn long-range global context and extract hierarchical feature maps from natural images. Based on this idea, SwinUNet [13] utilizes hierarchical Swin Transformer blocks to construct both encoder and decoder with a U-Net-like architecture. DS-TransUNet [78] adds on the more parallel encoder to process the input with a different resolution. SwinUNETR [120] uses a pre-training on a large medical image data set. Despite their usefulness, these fine-grain ViT-based approaches use standard self-attention to capture short- and long-range interactions. As such, they suffer from high computational costs and an explosion of time and memory costs, especially when the feature map size becomes large.

Potential of Hybrid Convolution and Transformer-based Operations. Accurately delineating boundaries in medical images in 3D multi-organ scenarios is a specific challenge, especially where tissues or organs overlap or have similar intensities [69]. Segmenting small or irregularly shaped structures demands high precision, which is crucial for many clinical applications. Additionally, inhomogeneities in tissue intensities within the same image pose another challenge [129], as do issues with image contrast and resolution [158], both of which can significantly impact segmentation accuracy. While CNN-based U-nets have demonstrated remarkable accuracy in medical image segmentation, they have limitations in modeling global dependencies due to localized receptive fields [87]. More recently, UNETR [48] and SwinUNETR [120] were proposed for a multi-organ/multi-tumor segmentation approach on 3D CT scans. These networks replace the CNN-based encoder with a Transformer or Swin Transformer in the U-Net and have achieved state-of-the-art accuracy [13, 48, 120]. However, it is worth noting that while Transformer is good at modeling global context, it is limited in capturing fine-grained details due to a lack of spatial inductive bias in modeling local information, especially for medical images. Although few works, such as TranFuse [148], try to solve this issue by presenting a fusion module to merge the features from CNN and Transformer, it is limited to 2D inputs. This leads to the question of whether a hybrid approach that combines the strengths of convolution and transformer-based operations could result in a more effective feature extraction encoder for medical image segmentation. By harnessing the local feature recognition capabilities of CNNs along with the global contextual understanding provided by transformers, such a hybrid model could offer a more comprehensive and nuanced analysis of medical images.

Label Scarcity and Annotation Challenges A significant challenge in medical image segmentation is the scarcity of labeled data. Annotating medical images requires expert knowledge and is often time-consuming and expensive [119]. This scarcity leads to models that may not generalize well across different datasets or clinical settings. The challenge becomes even more pronounced when dealing with multi-modal data such as CT and MRI images, which require separate annotations for each modality. The need for multi-modal annotations increases the workload for experts, further exacerbating the issue of label scarcity [105, 120]. Semi-supervised and unsupervised learning techniques [105, 120] have been proposed to mitigate this issue, leveraging unlabeled data to improve model performance. However, these methods still face challenges in ensuring high accuracy and reliability [130, 119].

One Framework for Multi-Modal Medical Image Segmentation In medical imaging and diagnostics, precise segmentation of anatomical structures from CT and MR imaging is vital for enhancing diagnostic and therapeutic practices. Humans can easily identify features across modalities, yet algorithms trained with single modalities struggle with multi-modal segmentation. Multi-modal learning, employing techniques like early fusion [55], late fusion [118], modality-specific representation [160], and hyperdense connections [29], improve segmentation by integrating information from diverse sources more effectively than single-modality methods. However, these traditional approaches require spatially aligned paired images from the same patient, a condition seldom met due to misalignments and variations in unpaired images [33], compromising performance. The shift towards unpaired multi-modal learning addresses this challenge by developing robust segmentation methods for unpaired images from different modalities, offering a practical, cost-effective solution for leveraging CT and MR in clinical settings.

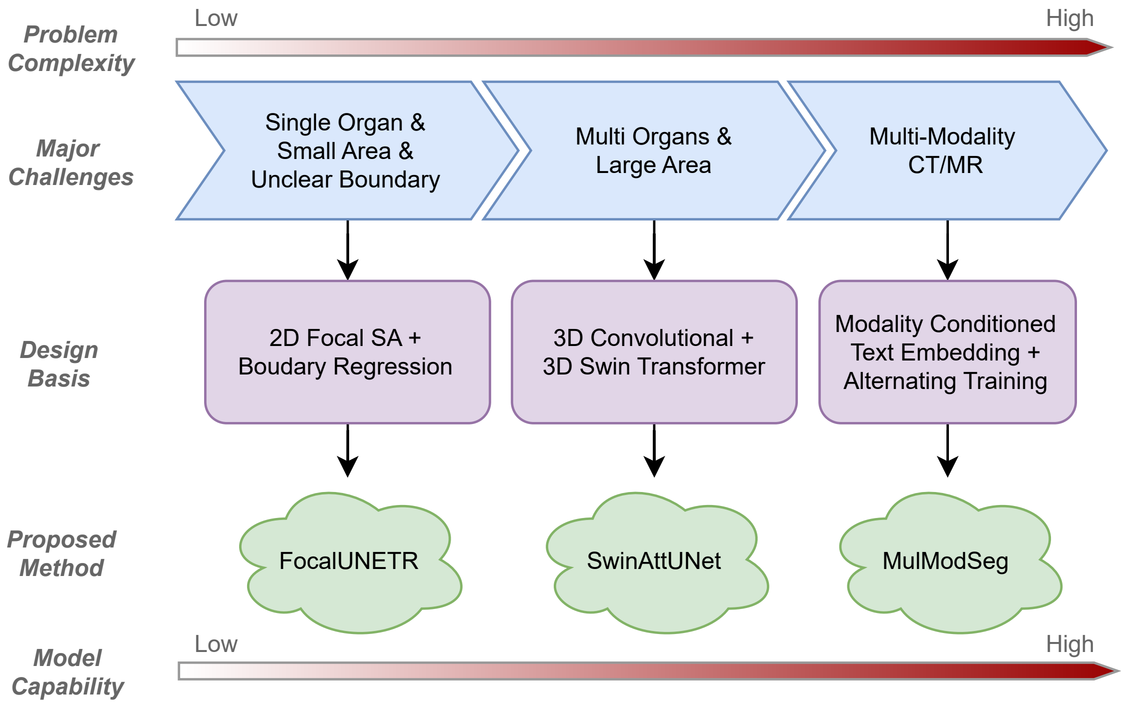
Problem Complexity
Low
High
Major Challenges
Single Organ & Small Area & Unclear Boundary
Multi Organs & Large Area
Multi-Modality CT/MR
Design Basis
2D Focal SA + Boudary Regression
3D Convolutional + 3D Swin Transformer
Modality Conditioned Text Embedding + Alternating Training
Proposed Method
FocalUNETR
SwinAttUNet
MulModSeg
Model Capability
Low
High

Figure 3: A schematic diagram elucidates the progression and interrelation among our three works: the 2D-based FocalUNETR, the 3D-based SwinAttUNet, and MulModSeg for multi-modal medical image segmentation. This hierarchy reflects an escalation in complexity, wherein the challenges associated with less complex problems are subsumed by those encountered in more intricate scenarios. Consequently, models with greater capabilities intrinsically encompass the functionalities of their less advanced counterparts.

1.3 Overview of Main Results

In this section, we will provide an overview of my research's core motivations, innovative ideas, and principal outcomes (as shown in Figure 3). In Section 1.3.1, it introduces FocalUNETR which incorporates focal self-attention to better leverage both local and global context for 2D CT-based medical image segmentation. Section 1.3.2, presents SwinAttUNet which utilizes both convolution and transformer-based modules for automatic 3D CT-based multi-organ segmentation. Lastly, Section 1.3.3, discusses MulModSeg, a multi-modal segmentation strategy that enhances medical image segmentation for both CT and MR images.

1.3.1 FocalUNETR: A Focal Transformer for Boundary-aware Segmentation of CT Images

Our proposed FocalUNETR [69] aims to address the limitations of CNNs in medical image segmentation, particularly for 2D prostate segmentation in CT scans. The motivation stems from the inherent challenges posed by unclear boundaries and high variability in such images. FocalUNETR introduces a transformer-based architecture that employs focal self-attention mechanisms, effectively enhancing the model's ability to process both local and global features. This novel approach aims to bridge the gap in accurately segmenting medical images with complex spatial relationships.

The research on FocalUNETR yields impactful and promising results, significantly outperforming current models in segmentation accuracy for prostate segmentation tasks on both private and public datasets. These advancements highlight its potential to enhance medical diagnostic precision and efficiency with effective and efficient focal self-attention mechanisms. Despite FocalUNETR's success, its application is currently limited to 2D single-organ segmentation. Due to the challenges in designing an efficient 3D version of the focal self-attention mechanism, a viable

3D multi-organ segmentation approach remains unachievable.

1.3.2 A New Architecture Combining Convolutional and Transformer-based Networks for Automatic 3D Multi-organ Segmentation on CT Images

The research presented in SwinAttUNet [70] introduces a novel architecture for 3D multi-organ segmentation in CT imaging, effectively addressing the limitations of existing deep learning models. This innovation stems from the understanding that while CNNs excel in recognizing local features, Transformers are adept at learning global, long-range context. The SwinAttUNet architecture merges these two approaches by combining a Swin Transformer with a CNN-based U-Net. This integration includes a parallel encoder to capture diverse features, a cross-fusion block for effective feature integration, and an attention-enabled decoder to enhance detail and context comprehension. This design is particularly adept at capturing both nuanced local details and broader global contextual information, making it a significant advancement in medical imaging technology.

SwinAttUNet notably outperforms other 3D-based state-of-the-art methods, both quantitatively and qualitatively. It shows statistically significant improvements in critical metrics like Dice Similarity Coefficients (DSC) or Dice scores, Hausdorff Distances (HD), and Average Surface Distances (ASD), thereby achieving superior accuracy in segmenting organs such as the prostate, bladder, rectum, lungs, liver, and kidneys. These advancements emphasize its potential to enhance the efficiency and consistency of medical image analysis, particularly in radiation therapy planning. However, it's important to acknowledge that SwinAttUNet's training involves a relatively small number of medical images. Given the abundance of labeled natural images, there is a pressing need to explore more effective ways to leverage these resources to further advance the field of medical image segmentation.

1.3.3 MulModSeg: Enhancing Unpaired Multi-Modal Medical Image Segmentation

We introduce MulModSeg, a multi-modal segmentation strategy designed to improve medical image segmentation for CT and MR modalities. Motivated by the challenges of varying imaging methods and data variability, MulModSeg incorporates modality-conditioned text embedding and an alternating training (ALT) procedure. The text embedding uses a pre-trained CLIP text encoder to integrate modality-specific information into the segmentation framework, enhancing feature extraction and accuracy. The ALT procedure alternates training between CT and MR images, ensuring balanced exposure and promoting model robustness across diverse medical imaging tasks.

Extensive experiments demonstrate that MulModSeg significantly outperforms state-of-the-art methods in segmentation accuracy for abdominal multi-organ and cardiac substructure tasks. Using both FCN and Transformer-based backbones, the research shows substantial improvements in critical metrics such as Dice scores. These results highlight MulModSeg's potential to enhance diagnostic accuracy and medical image analysis while seamlessly integrating into existing architectures without significant modifications. Overall, MulModSeg offers a robust and efficient solution for multi-modal medical image segmentation.

1.4 Organization of the Dissertation

The remainder of this dissertation is organized as follows: In Chapter 2, we present the FocalUNETR framework, which incorporates focal self-attention and a UNet-like multi-scale encoder-decoder design for 2D prostate segmentation. Additionally, we employ an auxiliary boundary-aware regression task during the training process to better address the issue of unclear boundaries in this specific task. In Chapter 3, we introduce a parallel convolutional and transformer-based 3D medical image encoder, resulting in the novel SwinAttUNet architecture for 3D multi-organ segmentation tasks. This hybrid design allows us to fully leverage the advantages of capturing both local spatial details and global long-range contexts. Chapter 4 introduces the MulModSeg method, which incorporates conditioned text embedding and alternating training techniques for CT/MR-based multi-modal medical image segmentation tasks. Finally, in Chapter 5, we summarize the main results and discuss future directions to follow based on the work we are currently working on.

1.5 Abbreviations

In Table 1, we summarize important abbreviations used throughout this dissertation.

Abbreviation	**Definition**
AG	attention gate
AI	artificial intelligence
AMOS	abdominal multi-Organ segmentation
ASD	average surface distance
ASSD	average symmetric surface distance
CNN	convolutional neural network
CLS	class
CT	computed tomography
CT-ORG	CT organ segmentation dataset
DL	deep learning
DSC	dice similarity coefficient
FCN	fully convolutional network
HD	Hausdorff distance
HU	Hounsfield unit
MLP	multi-layer perception
MRI	magnetic resonance imaging
MR	magnetic resonance
NLP	natural language processing
NN	neural network
SOTA	state of the art
ViT	vision transformer

Table 1: A summary of important abbreviations used throughout the dissertation.

CHAPTER 2 FOCALUNETR: A FOCAL TRANSFORMER FOR BOUNDARY-AWARE SEGMENTATION OF CT IMAGES

2.1 Introduction

2.1.1 Background Significance

Prostate cancer is a leading cause of cancer-related deaths in adult males, as highlighted in studies such as Parikesit et al. [104]. A common treatment option for prostate cancer is external beam radiation therapy [28], where CT scanning serves as a cost-effective tool for treatment planning compared to the more expensive MRI. Precise prostate segmentation in CT images is crucial to ensure effective radiation dose delivery to tumor tissues while minimizing harm to surrounding healthy tissues.

However, CT images have relatively low spatial resolution and soft tissue contrast compared to MRI, making manual prostate segmentation time-consuming and prone to significant inter-operator variability [76]. This poses a significant challenge in the treatment planning process, necessitating the development of automated segmentation methods that can consistently deliver accurate results.

2.1.2 Related Work

Several automated segmentation methods, particularly those based on fully convolutional networks (FCNs) such as U-Net [110] and its variants [95, 136, 164], have been proposed to address the limitations of manual segmentation. Despite their progress, these methods often struggle to capture long-range relationships and global context information [20] due to the inherent limitations of convolutional operations.

To overcome these limitations, researchers have turned to Vision Transformers (ViT) [30], which leverage self-attention (SA) mechanisms. TransUNet [20] was one of the first to adapt ViT for medical image segmentation by integrating transformer layers with an FCN-based encoder to better capture global context from high-level feature maps. Other models, such as TransFuse [148] and MedT [123], combine FCNs and Transformers to capture both global dependencies and low-level spatial details more effectively. Swin-UNet [13], utilizing efficient Swin Transformers [85], and models like UNETR [48] and SwinUNETR [120] extended for 3D inputs, have also shown promising results.

Despite these advancements, ViT-based networks using standard or shifted-window SA often overlook the interactions between local and global contexts [138, 107]. As reported by Tang et al. [120], even with self-supervised pre-training on extensive medical data, the performance of prostate segmentation in high-resolution MRI images remains unsatisfactory, let alone in lower-quality CT images. Additionally, the unclear prostate boundaries in CT images due to low soft tissue contrast pose significant challenges [50, 131].

2.1.3 Our Contribution

Recently, the Focal Transformer [138] was proposed for general computer vision tasks, introducing focal self-attention to incorporate both fine-grained local and coarse-grained global interactions. Inspired by this work, we propose FocalUNETR (Focal U-NEt TRansformers), a novel architecture for CT-based medical image segmentation (Fig. 4A). While previous works like Psi-Net [99] have incorporated additional decoders for enhanced boundary detection, they often fail to capture global context effectively or address boundary randomness in low-contrast CT images. In contrast, our approach employs a multi-task learning strategy with a Gaussian kernel over the ground truth segmentation mask boundary [79] as an auxiliary boundary-aware contour regression task (Fig. 4B). This auxiliary task acts as a regularization term for the main segmentation task, enhancing the model's ability to handle unclear boundaries in CT images.

In this chapter, we make several key contributions. First, we develop a novel focal transformer model (FocalUNETR) for CT-based prostate segmentation, utilizing focal SA to hierarchically learn feature maps that account for both short- and long-range visual dependencies. We also tackle the challenge of unclear boundaries specific to CT images by incorporating an auxiliary contour regression task. Our methodology demonstrates superior performance compared to state-of-the-art methods through extensive experiments on both private and public datasets.

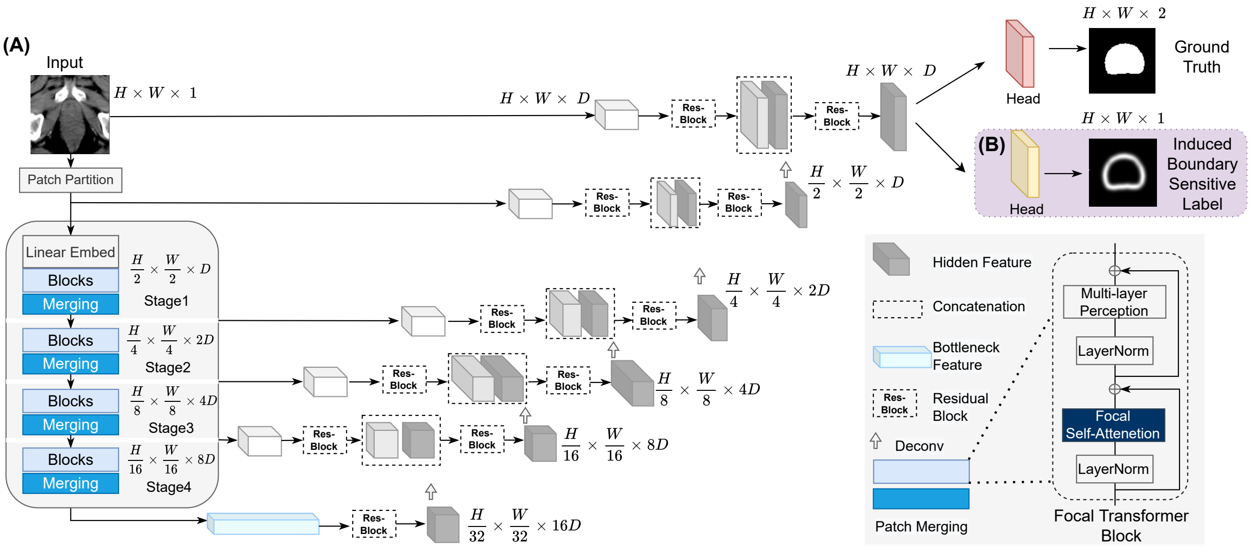

Figure 4: The architecture of FocalUNETR (A) as the main task for prostate segmentation and a boundary-aware regression auxiliary task (B).

2.2 The FocalUNETR Architecture

2.2.1 The Main Task for Mask Generation

Our FocalUNETR architecture (Fig. 4) follows a multi-scale design similar to [48, 120], enabling us to obtain hierarchical feature maps at different stages. The input medical image $\mathcal{X} \in \mathcal{R}^{C \times H \times W}$ is first split into a sequence of tokens with dimension $\lceil \frac{H}{H'} \rceil \times \lceil \frac{W}{W'} \rceil$, where H, W represent spatial height and width, respectively, and C represents the number of channels. These tokens are then projected into an embedding space of dimension D using a patch of resolution

(H', W'). The SA is computed at two focal levels [138]: fine-grained and coarse-grained, as illustrated in Fig. 5A. The focal SA attends to fine-grained tokens locally, while summarized tokens are attended to globally (reducing computational cost). We perform focal SA at the window level, where a feature map of $x \in \mathcal{R}^{d \times H'' \times W''}$ with spatial size $H'' \times W''$ and d channels is partitioned into a grid of windows with size $s_w \times s_w$. For each window, we extract its surroundings using focal SA.

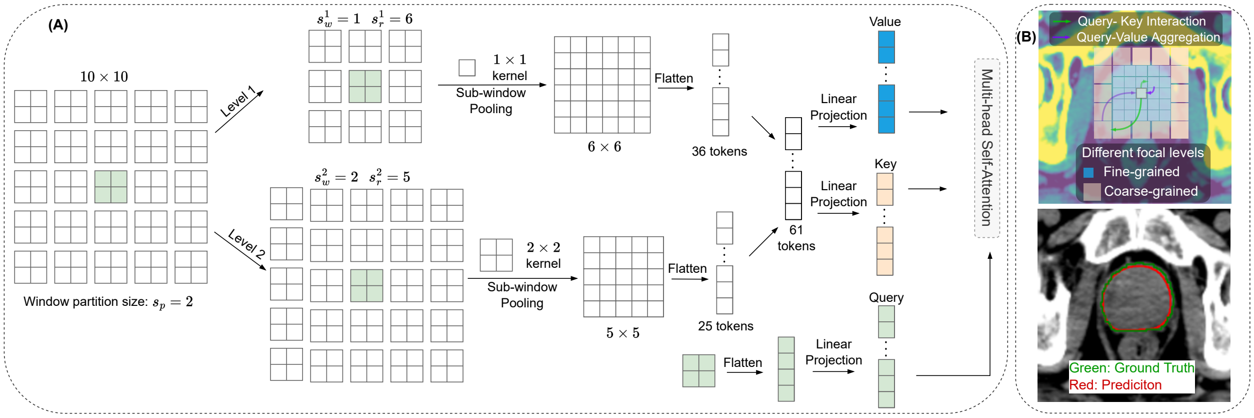

Figure 5: (A) The focal SA mechanism, and (B) an example of perfect boundary matching using focal SA for CT-based prostate segmentation task (lower panel), in which focal SA performs query-key interactions and query-value aggregations in both fine- and coarse-grained levels (upper panel).

For window-wise focal SA [138], there are three terms $\{L, s_w, s_r\}$. Focal level L is the number of granularity levels for which we extract the tokens for our focal SA. We present an example, depicted in Fig. 5B, that illustrates the use of two focal levels (fine and coarse) for capturing the interaction of local and global context for optimal boundary-matching between the prediction and the ground truth for prostate segmentation. Focal window size s_w^l is the size of the sub-window on which we get the summarized tokens at level $l \in \{1, \ldots, L\}$. Focal region size s_r^l is the number of sub-windows horizontally and vertically in attended regions at level l. The focal SA module proceeds in two main steps, sub-window pooling and attention computation. In the sub-window pooling step, an input feature map $x \in \mathcal{R}^{d \times H'' \times W''}$ is split into a grid of sub-windows with size

$\{s_w^l, s_w^l\}$, followed by a simple linear layer f_p^l to pool the sub-windows spatially. The pooled feature maps at different levels l provide rich information at both fine-grained and coarse-grained, where $x^l = f_p^l(\hat{x}) \in \mathcal{R}^{d \times \frac{H''}{s_w^l} \times \frac{W''}{s_w^l}}$, and $\hat{x} = \text{Reshape}(x) \in \mathcal{R}^{(d \times \frac{H''}{s_w^l} \times \frac{W''}{s_w^l}) \times (s_w^l \times s_l^w)}$. After obtaining the pooled feature maps $x^{l}{}_1^L$, we calculate the query at the first level and key and value for all levels using three linear projection layers f_q, f_k, and f_v:

$$Q = f_q(x^1), K = \{K^l\}_1^L = f_k(\{x^1, \ldots, x^L\}), V = \{V^l\}_1^L = f_v(\{x^1, \ldots, x^L\}). \tag{2.1}$$

For the queries inside the i-th window $Q_i \in \mathcal{R}^{d \times s_w \times s_w}$, we extract the $s_r^l \times s_r^l$ keys and values from K^l and V^l around the window where the query lies in and then gather the keys and values from all L to obtain $K_i = \{K_1, \ldots, K_L\} \in \mathcal{R}^{s \times d}$ and $V_i = \{V_1, \ldots, V_L\} \in \mathcal{R}^{s \times d}$, where $s = \sum_{l=1}^{L} (s_r^l)^2$. Finally, a relative position bias is added to compute the focal SA for Q_i by

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + B) V_i, \tag{2.2}$$

where $B = \{B^l\}_1^L$ is the learnable relative position bias [138].

The encoder utilizes a patch size of 2 × 2 with a feature dimension of $2 \times 2 \times 1 = 4$ (i.e., a single input channel CT) and a D-dimensional embedding space. The overall architecture of the encoder comprises four stages of focal transformer blocks, with a patch merging layer applied between each stage to reduce the resolution by a factor of 2. We utilize an FCN-based decoder (Fig. 4A) with skip connections to connect to the encoder at each resolution to construct a “U-shaped" architecture for our CT-based prostate segmentation task. The output of the encoder is concatenated with processed input volume features and fed into a residual block. A final 1 × 1 convolutional layer with a suitable activation function, such as Softmax, is applied to obtain the required number of class-based probabilities.

For the experiments, we follow the hyperparameter settings suggested in [138], with 2 focal levels, transformer blocks of depths [2, 2, 6, 2], and head numbers [4, 8, 16, 32] for each of the four stages. We then create FocalUNETR-S and FocalUNETR-B with D as 48 and 64, respectively. These settings have parameters of 27.3 M and 48.3 M, which are comparable to other state-of-the-art models.

2.2.2 The Auxiliary Task for Boundary Regression

For the main task of mask prediction (as illustrated in Fig. 4A), a combination of Dice loss and Cross-Entropy loss is employed to evaluate the concordance of the predicted mask and the ground truth on a pixel-wise level. The objective function for the segmentation head is given by: $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{dice}}(\hat{p}_i, G) + \mathcal{L}_{\text{ce}}(\hat{p}_i, G)$, where $\hat{p}_i$ represents the predicted probabilities from the main task and G represents the ground truth mask, both given an input image i. The predicted probabilities, $\hat{p}_i$, are derived from the main task through the application of the FocalUNETR model to the input CT image.

To address the challenge of unclear boundaries in CT-based prostate segmentation, an auxiliary task is introduced for the purpose of predicting boundary-aware contours to assist the main prostate segmentation task. This auxiliary task is achieved by attaching another convolution head after the extracted feature maps at the final stage (see Fig. 4B). The boundary-aware contour, or the induced boundary-sensitive label, is generated by considering pixels near the boundary of the prostate mask. To do this, the contour points and their surrounding pixels are formulated into a Gaussian distribution using a kernel with a fixed standard deviation of σ (in this specific case, e.g., $\sigma = 1.6$) [89, 50, 79]. The resulting contour is a heatmap in the form of a *Heatsum* function [79].

We predict this heatmap with a regression task trained by minimizing mean-squared error instead of treating it as a single-pixel boundary segmentation problem. Given the ground truth of contour G_i^C, induced from the segmentation mask for input image i, and the reconstructed output probability $\hat{p}_i^C$, we use the following loss function: $\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_i \|\hat{p}_i^C - G_i^C\|_2$ where N is the total number of images for each batch. This auxiliary task is trained concurrently with the main segmentation task.

A multi-task learning approach is adopted to regularize the main segmentation task through the auxiliary boundary prediction task. The overall loss function is a combination of $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{reg}}$: $\mathcal{L}_{\text{tol}} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{reg}}$, where λ_1 and λ_2 are hyper-parameters that weigh the contribution of the mask prediction loss and contour regression loss, respectively, to the overall loss. The optimal setting of $\lambda_1 = \lambda_2 = 0.5$ is determined by trying different settings.

2.3 Experiments

2.3.1 Datasets

To evaluate our method, we use a large private dataset with 400 CT scans and a large public dataset with 300 CT scans (AMOS [57]). As far as we know, the AMOS dataset is the only publicly available CT dataset including prostate ground truth. We randomly split the private dataset with 280 scans for training, 40 for validation, and 80 for testing. The AMOS dataset has 200 scans for training and 100 for testing [57]. Although the AMOS dataset includes the prostate class, it mixes the prostate (in males) and the uterus (in females) into one single class labeled PRO/UTE. We filter out CT scans missing the PRO/UTE ground-truth segmentation.

2.3.2 Implementation Details

For the implementation, we utilize a server equipped with 8 Nvidia A100 GPUs, each with 40 GB of memory. All experiments are conducted in PyTorch, and each model is trained on a single GPU. We interpolate all CT scans into a constant voxel spacing of $[1.0 \times 1.0 \times 1.5]$ *mm* for both datasets. Houndsfield unit (HU) range of $[-50, 150]$ is used and normalized to $[0, 1]$. Subsequently, each CT scan is cropped to a $128 \times 128 \times 64$ voxel patch around the prostate area, which is used as input for 3D models. For 2D models, we first slice each voxel patch in the axial direction into 64 slices of 128×128 images for training and stack them back for evaluation. For the private dataset, we train models for 200 epochs using the AdamW optimizer with an initial learning rate of $5e^{-4}$. An exponential learning rate scheduler with a warmup of 5 epochs is applied to the optimizer. The batch size is set to 24 for 2D models and 1 for 3D models. We use random flip, rotation, and intensity scaling as augmentation transforms with probabilities of 0.1, 0.1, and 0.2, respectively. We also tried using 10% percent of AMOS training set as validation data to find a better training parameter setting and re-trained the model with the full training set. However, we did not get improved performance compared with directly applying the training parameters learned from tuning the private dataset. We report the Dice Similarity Coefficient (DSC, %), 95% percentile Hausdorff Distance (HD, mm), and Average Symmetric Surface Distance (ASSD, mm) metrics.

2.4 Results and Discussion

2.4.1 Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of FocalUNETR, we compare the CT-based prostate segmentation performance with three 2D U-Net-based methods: U-Net[110], UNet++ [164], and Attention U-Net (AttUNet) [101], two 2D transformer-based segmentation methods: TransUNet [20] and Swin-UNet [13], two 3D U-Net-based methods: U-Net (3D) [23] and V-Net [95], and two 3D transformer-based models: UNETR [48] and SiwnUNETR [120]. nnUNet [56] is used for

Method	Private			AMOS		
	DSC ↑	HD ↓	ASSD ↓	DSC ↑	HD ↓	ASSD ↓
U-Net	85.22 (1.23)	6.71 (1.03)	2.42 (0.65)	83.42 (2.28)	8.51 (1.56)	2.79 (0.61)
UNet++	85.53 (1.61)	6.52 (1.13)	2.32 (0.58)	83.51 (2.31)	8.47 (1.62)	2.81 (0.57)
AttUNet	85.61 (0.98)	6.57 (0.96)	2.35 (0.72)	83.47 (2.34)	8.43 (1.85)	2.83 (0.59)
TransUNet	85.75 (2.01)	6.43 (1.28)	2.23 (0.67)	81.13 (3.03)	9.32 (1.87)	3.71 (0.79)
Swin-UNet	86.25 (1.69)	6.29 (1.31)	2.15 (0.51)	83.35 (2.46)	8.61 (1.82)	3.20 (0.64)
U-Net (3D)	85.42 (1.34)	6.73 (0.93)	2.36 (0.67)	83.25 (2.37)	8.43 (1.65)	2.86 (0.56)
V-Net (3D)	84.42 (1.21)	6.65 (1.17)	2.46 (0.61)	81.02 (3.11)	9.01 (1.93)	3.76 (0.82)
UNETR (3D)	82.21 (1.35)	7.25 (1.47)	2.64 (0.75)	81.09 (3.02)	8.91 (1.86)	3.62 (0.79)
SwinUNETR (3D)	84.93 (1.26)	6.85 (1.21)	2.48 (0.52)	83.32 (2.23)	8.63 (1.62)	3.21 (0.68)
nnUNet	85.86 (1.31)	6.43 (0.91)	2.09 (0.53)	83.56 (2.25)	8.36 (1.77)	***2.65 (0.61)***
FocalUNETR-S	86.53 (1.65)	5.95 (1.29)	2.13 (0.29)	82.21 (2.67)	8.73 (1.73)	3.46 (0.75)
FocalUNETR-B	*87.73 (1.36)*	*5.61 (1.18)*	*2.04 (0.23)*	*83.61 (2.18)*	*8.32 (1.53)*	2.76 (0.69)
FocalUNETR-S*	87.84 (1.32)	5.59 (1.23)	2.12 (0.31)	83.24 (2.52)	8.57 (1.70)	3.04 (0.67)
FocalUNETR-B*	**89.23 (1.16)**	**4.85 (1.05)**	**1.81 (0.21)**	**83.79 (1.97)**	**8.31 (1.45)**	2.71 (0.62)

Table 2: Quantitative performance comparison on the private and AMOS datasets with a mean (standard deviation) for 3 runs with different seeds. An asterisk (*) denotes the model is co-trained with the auxiliary contour regression task. The best results with/without the auxiliary task are boldfaced or italicized, respectively.

comparison as well. Both 2D and 3D models are included as there is no conclusive evidence for which type is better for this task [131]. All methods (except nnUNet) follow the same settings as FocalUNETR and are trained from scratch. TransUNet and Swin-UNet are the only methods that are pre-trained on ImageNet. Detailed information regarding the number of parameters, FLOPs, and average inference time can be found in Table 3.

Quantitative results are presented in Table 2, which shows that the proposed FocalUNETR, even without co-training, outperforms other FCN and Transformer baselines (2D and 3D) in both datasets for most of the metrics. The AMOS dataset mixes the prostate (males) /uterus (females, a relatively small portion). The morphology of the prostate and uterus is significantly different. Consequently, the models may struggle to provide accurate predictions for this specific portion of the uterus. Thus, the overall performance of FocalUNETR is overshadowed by this challenge, resulting in only moderate improvement over the baselines on the AMOS dataset. However, the performance in the real-world (private) dataset gains a much better performance margin. When co-trained with the auxiliary contour regression task and using the multi-task training strategy, the performance of FocalUNETRs is further improved. In summary, these observations indicate that incorporating FocalUNETR and multi-task training with an auxiliary contour regression task can improve the challenging CT-based prostate segmentation performance.

Qualitative results of several representative methods are visualized in Fig. 6. The figure shows that our FocalUNETR-B and FocalUNETR-B* generate more accurate segmentation results that are more consistent with the ground truth than the results of the baseline models. All methods perform well for relatively easy cases (1^{st} row in Fig. 6), but the FocalUNETRs outperform the other methods. For more challenging cases (rows 2-4 in Fig. 6), such as unclear boundaries and mixed PRO/UTE labels, FocalUNETRs still perform better than other methods. Additionally, the FocalUNETRs are less likely to produce false positives (as shown in Fig. 7) for CT images without a foreground ground truth, due to the focal SA mechanism that enables the model to capture global context and helps to identify the correct boundary and shape of the prostate. Overall, the FocalUNETRs demonstrate improved segmentation capabilities while preserving shapes more precisely, making them promising tools for clinical applications.

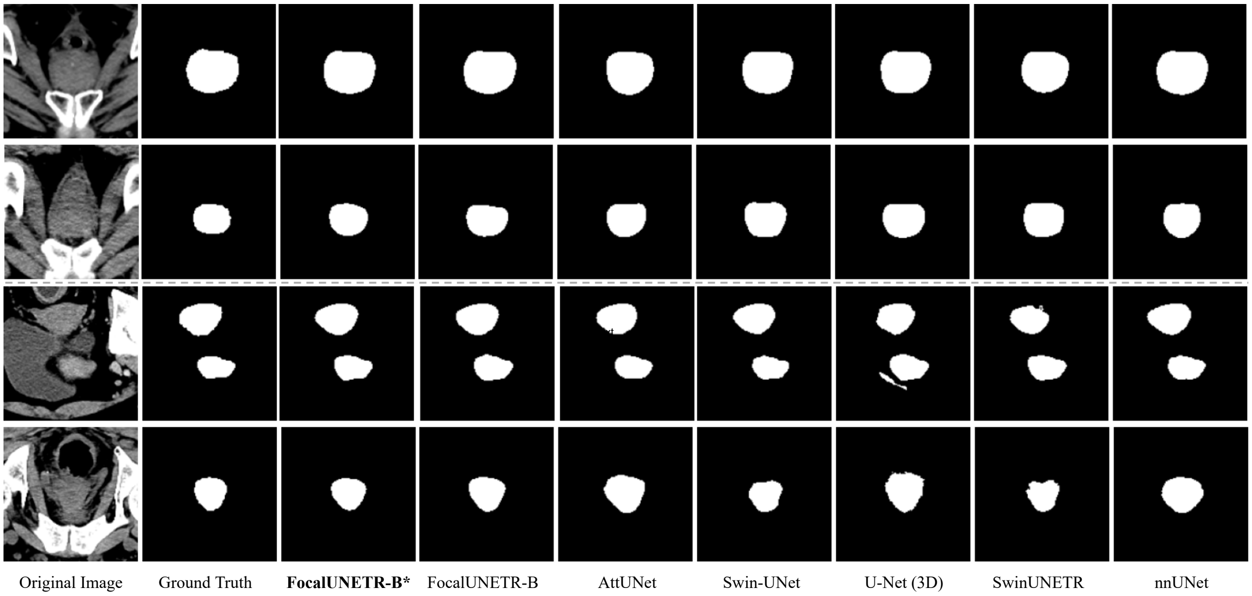

Figure 6: Qualitative results on sample test CT images from the private (first two rows) and AMOS (last two rows) datasets.

2.4.2 Parameters and Inference Time

As shown in Table 3, our proposed FocalUNETR demonstrates a comparable model size, relatively small FLOPs, and fast inference speed to most of the SOTAs. The parameter count for FocalUNETR-S is 27.3 million with 15.7 GFLOPs, and for FocalUNETR-B, it is 48.3 million with 27.5 GFLOPs. This positions FocalUNETR as a moderate-sized model with efficient computational requirements. For instance, models like TransUNet and UNETR (3D) have significantly larger parameter counts of 105.3 million and 92.6 million, and FLOPs of 29.3G and 75.4G, re-

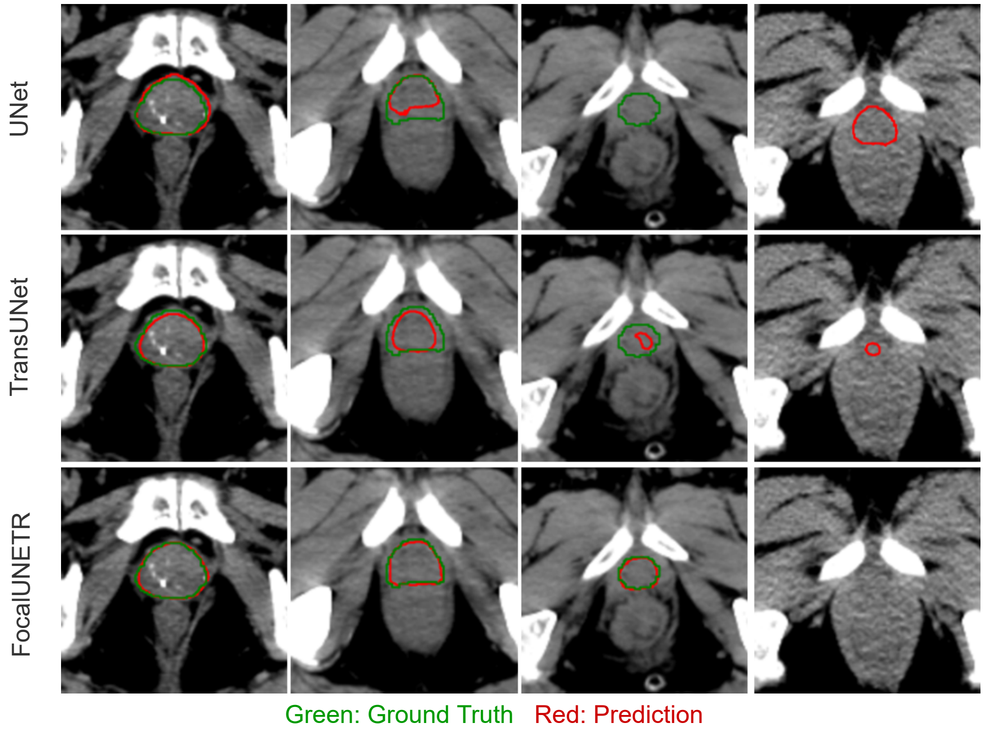

Figure 7: Qualitative results of prostate segmentation by comparing our FocalUNETR-B with UNet and TransUNet in 2D settings. All methods perform well for easy cases, but our FocalUNETR-B can be even better. FocalUNETR-B is less likely to give a false prediction (false positives) for CT images without a foreground mask.

spectively, resulting in longer inference times of 4.87s and 6.49s. In contrast, FocalUNETR-S achieves an inference time of 4.36s, and FocalUNETR-B achieves 5.35s, which is competitive with models such as Swin-UNet (3.58s) and U-Net (3.12s) while maintaining lower FLOPs compared to models like nnUNet (389G) and SwinUNETR (3D) (350G). These characteristics highlight FocalUNETR's balance between model size, computational efficiency, and speed, making it an effective choice for medical image segmentation tasks.

Model	Param. (M)	FLOPs (G)	Average Inference Time (s)
U-Net	7.2	9.3	3.12
UNet ++	22.5	60.4	4.31
AttUNet	19.8	25.5	3.53
TransUNet	105.3	29.3	4.87
Swin-UNet	41.4	9.0	3.58
U-Net (3D)	16.6	285	6.51
V-Net (3D)	45.6	586	6.72
UNETR (3D)	92.6	75.4	6.49
SwinUNETR (3D)	62.2	350	7.23
nnUNet	19.3	389	9.65
FocalUNETR-S	27.3	15.7	4.36
FocalUNETR-B	48.3	27.5	5.35

Table 3: The number parameters, FLOPs, and average inference time per case for different models: our FocalUNETR shows a comparable model size, relatively small FLOPs, and fast inference speed to most of the SOTAs.

2.4.3 Ablation Study

To better examine the efficacy of the auxiliary task for FocalUNETR, we selected different settings of λ_1 and λ_2 for the overall loss function $\mathcal{L}_{tol}$ on the private dataset. The results (Table 4) indicate that as the value of λ_2 is gradually increased and that of λ_1 is correspondingly decreased (thereby increasing the relative importance of the auxiliary contour regression task), segmentation performance initially improves. However, as the ratio of contour information to segmentation mask information becomes too unbalanced, performance begins to decline. Thus, it can be inferred that the optimal setting for these parameters is when both λ_1 and λ_2 are set to 0.5.

$\mathcal{L}_{\text{tol}}$	$\mathcal{L}_{\text{seg}}$	$0.8\mathcal{L}_{\text{seg}} + 0.2\mathcal{L}_{\text{reg}}$	$0.5\mathcal{L}_{\text{seg}} + 0.5\mathcal{L}_{\text{reg}}$	$0.2\mathcal{L}_{\text{seg}} + 0.8\mathcal{L}_{\text{reg}}$
DSC ↑	87.73 ± 1.36	88.01 ± 1.38	**89.23 ± 1.16**	87.53 ± 2.13

Table 4: Ablation study on different settings of total loss for FocalUNETR-B on the private dataset.

2.5 Conclusion

The proposed FocalUNETR architecture offers a transformative solution for precise prostate segmentation in CT imaging, addressing challenges such as the prostate's unclear boundaries due to poor soft tissue contrast and the limitations of convolutional neural network-based models in capturing long-range global context. This novel focal transformer-based image segmentation architecture efficiently extracts both local visual features and global context from CT images. An innovative addition to this approach is the auxiliary boundary-induced label regression task, specifically designed to tackle the issue of unclear boundaries in low-contrast CT images. The effectiveness of FocalUNETR is evident in its substantial improvements in the Dice Similarity Coefficient, reduced Hausdorff Distance, and Average Symmetric Surface Distance, outperforming other methods on both private and public CT datasets. Despite its success, the architecture currently focuses only on prostate segmentation and requires further development for 3D input adaptation and multi-organ segmentation.

While FocalUNETR has achieved notable success, its application is presently confined to 2D-based single-organ segmentation. The difficulty in developing an efficient 3D version of the focal SA mechanism poses a significant challenge, leaving the realization of a 3D-based multi-organ segmentation approach out of reach. To address this issue, the forthcoming Chapter 3 introduces an innovative architecture. This new design uniquely combines convolutional and transformer-based operations in parallel during the feature extraction stage, specifically tailored for the more complex 3D-based multi-organ segmentation tasks.

CHAPTER 3 A NEW ARCHITECTURE COMBINING CONVOLUTIONAL AND TRANSFORMER-BASED NETWORKS FOR AUTOMATIC 3D MULTI-ORGAN SEGMENTATION ON CT IMAGES

3.1 Introduction

3.1.1 Background Significance

In the field of radiation therapy, precise targeting of tumor tissue while avoiding normal tissues is crucial for successful treatment [10, 121, 143]. One of the key steps in the planning process involves segmenting the treatment target and organs-at-risk (OARs) typically using planning CT images. Currently, the clinical practice for contour delineation involves a labor-intensive and operator-dependent manual process [46, 36, 39]. The manual contouring process in addition to often being inefficient can also suffer from inconsistencies in contouring preferences or related intra-and inter-observer uncertainties [46, 39]. Inaccuracies in contouring impact on planning margin design—erroneous planning margins may lead to possible underdosing of the target and excess radiation delivered to surrounding healthy tissues [127]. To address these issues, a method for accurate automatic segmentation is needed to improve efficiency and consistency in radiation treatment planning.

3.1.2 Related Work

Modern automatic multi-organ segmentation models can be roughly classified into two categories: conventional learning and deep learning-based segmentation [143, 44, 77, 11]. In general, conventional learning-based approaches for building segmentation models have two major components [54] : (a) extraction of handcrafted features to represent target organs, and (b) classification/regression model for segmentation. For instance, Glocker et al. [43] developed a supervised forest model that uses both class and structural information to jointly perform pixel classification and shape regression. To enhance the segmentation performance, Chen and Zheng [19] selected the most important features from the complete feature set using a hierarchical landmark detection method. Gao et al.[40] utilized multi-task random forests to segment the prostate, bladder, rectum, and left and right femoral heads, jointly with a displacement regression task. Since these methods are typically created using low-dimensional hand-crafted features, their performance may be limited, particularly when the training datasets suffer from limited contrast impeding clear differentiation between organs at the boundaries, as is sometimes encountered with CT images.

Recently deep learning algorithms, which rely primarily on fully convolutional neural networks (CNNs) based U-net architectures [110, 95, 62, 55, 15, 26, 116, 141] have been applied to the problem of organ segmentation for radiation treatment planning [143, 142, 113]. The U-Net is a popular architecture and comprises an encoder and decoder, where the encoder progressively reduces the resolution of CT scans to generate conceptual features across multiple scales. The decoder then reconstructs the extracted features for multi-organ segmentation. The U-net model incorporates skip-connections that combine the encoder and decoder outputs at different resolutions to maintain information lost during downsampling and improve performance. In pelvic organ segmentation, advanced U-net algorithms utilize supplementary techniques to facilitate the learning of more informative segmentation features. These techniques include a localization network for detecting the location of each organ before pixel-level segmentation [5], a self-attention/Transformer mechanism for acquiring global features [102], deep supervision for improving generality [63], and multi-task learning strategies for capturing boundaries [131].

While CNN-based U-Nets have demonstrated promise for medical image segmentation, they have limitations in modeling global context (as shown in Figure 1) because the learning approach tends to be focused on local information [87]. To overcome this limitation, the vision Transformer

(ViT) [30] has been proposed as an effective method to capture global dependencies and improve segmentation results for object structures with varying sizes and shapes. Studies have explored the integration of Transformers into U-Net architectures to enhance their performance in CT image segmentation. For instance, Chen et al. [20] used a Transformer between the encoder and decoder of U-Net to segment 2D abdominal CT scans and capture global context from U-Net feature maps. Similarly, Cao et al. [14] proposed a U-Net with a shifted-window (Swin) transformer (Swin-Unet) for 2D CT/MRI segmentation by replacing the convolutional blocks in U-Net with Swin Transformer blocks for both the encoder and decoder. More recently, UNETR [48] and SwinUNETR [120] were proposed for a multi-organ/multi-tumor segmentation approach on 3D CT scans. These networks replace the CNN-based encoder with a Transformer or Swin Transformer in the U-Net and have achieved state-of-the-art accuracy [14, 48, 120, 148]. However, it is worth noting that while Transformer is effective at modeling global context, it is limited in capturing granular details due to a lack of spatial inductive bias in modeling local information, especially in the low data (high background) setting as is encountered with medical images [140, 49].

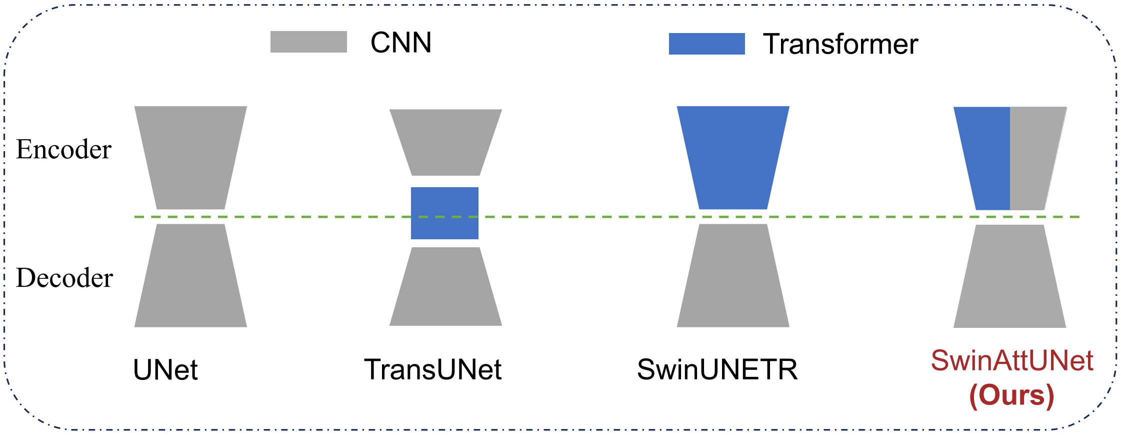

Figure 8: A comparison of several architectures for medical image segmentation with the encoder-decoder architecture.

3.1.3 Our Contribution

In this chapter, we developed and optimized a novel architecture, termed "SwinAttUNet" for 3D CT-based auto-segmentation of the prostate gland and surrounding OARs, and other normal organs, including the lungs, liver, kidneys, and pelvic bones. SwinAttUNet bridges a 3D-U-Net and a Swin Transformer in a parallel encoding manner (as shown in Figure 8) to take advantage of both architectures. SwinAttUNet includes a parallel encoder, a cross-fusion block, and a CNN-based decoder. To our knowledge, this is the first network combining a 3D-based parallel CNN with a Transformer, along with several other unique features, for multiple organ segmentation. Details of the network architecture and quantitative evaluation of the model are presented.

3.2 The SwinAttUNet Architecture

3.2.1 Overall Architecture Design

As depicted in Figure 9, we introduce the SwinAttUNet, which is a 3D network and is trained using 3D CT image datasets. SwinAttUNet includes a parallel encoder, a cross-fusion block, and a CNN-based decoder. The parallel encoder consists of a CNN branch (CB) and a Transformer branch (TB), which independently extracts local details and global contextual information. The cross-fusion block merges local and global features on the same scale. The CNN-based decoder is designed to adapt the fused information and thereby improve model stability while maintaining performance. A skip connection is applied between the cross-fusion block and decoder to integrate low-level semantic features. Attention gates (AGs) are integrated within the CNN to suppress features in image background regions and focus attention on important regions of the image (targets and OARs). All convolution blocks are 3D convolutions with a kernel size of three, and all transformer blocks are Swin Transformers (with window-based self-attention and shifted-window-

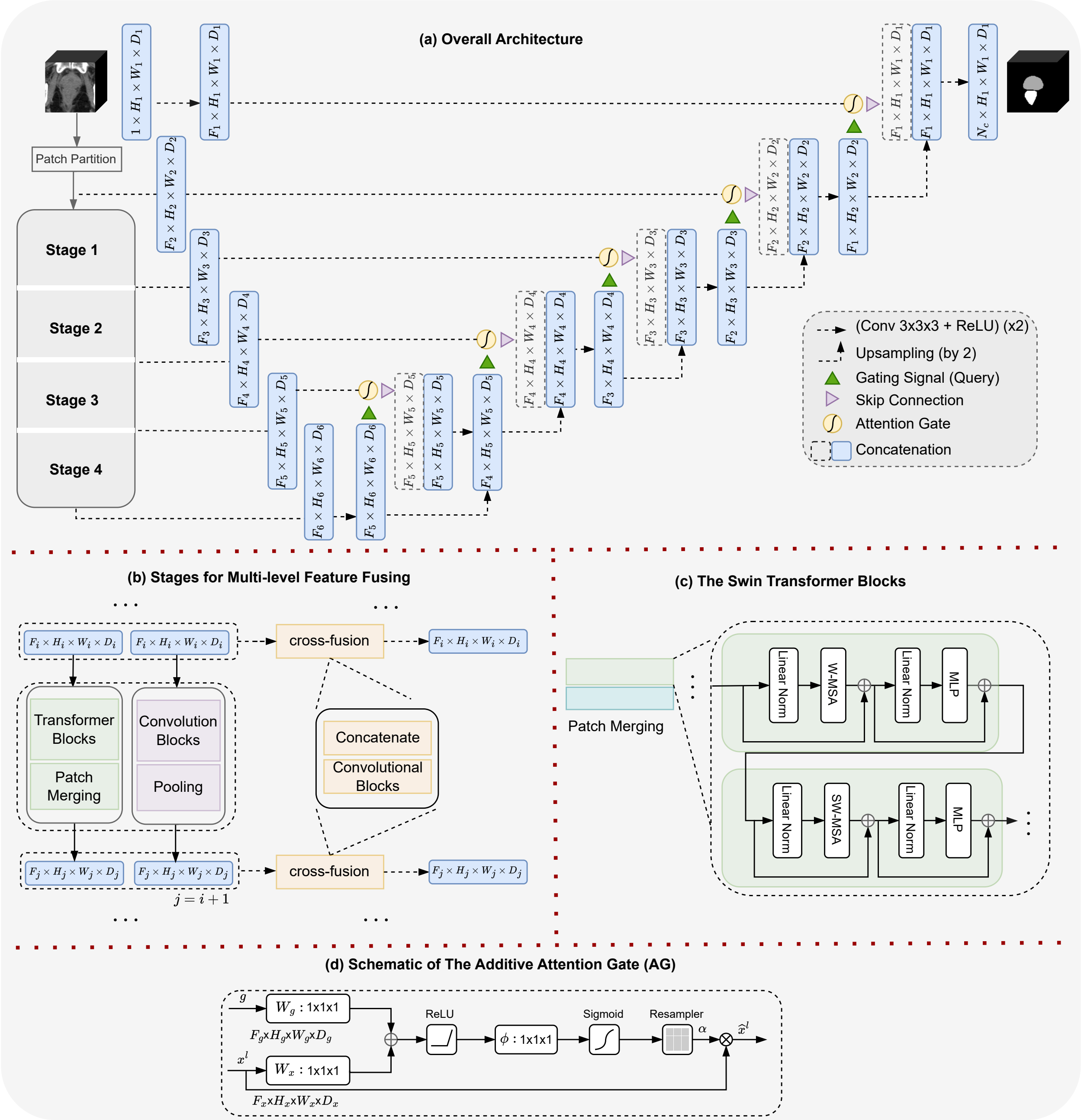
(a) Overall Architecture
Patch Partition
Stage 1
Stage 2
Stage 3
Stage 4
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_2 \times H_2 \times W_2 \times D_2$
$F_3 \times H_3 \times W_3 \times D_3$
$F_4 \times H_4 \times W_4 \times D_4$
$F_5 \times H_5 \times W_5 \times D_5$
$F_6 \times H_6 \times W_6 \times D_6$
$N_c \times H_1 \times W_1 \times D_1$
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Gating Signal (Query)
Skip Connection
Attention Gate
Concatenation
(b) Stages for Multi-level Feature Fusing
$F_i \times H_i \times W_i \times D_i$
cross-fusion
Transformer Blocks
Patch Merging
Convolution Blocks
Pooling
Concatenate
Convolutional Blocks
$F_j \times H_j \times W_j \times D_j$
$j = i + 1$
(c) The Swin Transformer Blocks
Patch Merging
Linear Norm
W-MSA
MLP
SW-MSA
(d) Schematic of The Additive Attention Gate (AG)
g
W_g : 1x1x1
$F_g x H_g x W_g x D_g$
x^l
W_x : 1x1x1
$F_x x H_x x W_x x D_x$
ReLU
ϕ : 1x1x1
Sigmoid
Resampler
α
$\hat{x}^l$

Figure 9: (a) The architecture of SwinAttUNet for pelvic segmentation with 3D CT inputs. (b) Parallel CNN and Transformer blocks for encoder with a cross-fusion module. (c) The architecture of two successive Swin Transformer Blocks, W-MSA, and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively. (d) Schematic of the proposed additive attention gate (AG). Input features (x^l) are scaled with attention coefficients (α) computed in AG. Spatial regions are selected by analyzing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale. Grid resampling of attention coefficients is done using trilinear interpolation.

based self-attention). We use two blocks for both convolution and transformer operations.

3.2.2 Swin Transformer Branch for 3D Inputs

Our SwinAttUNet architecture features a multi-scale design that enables the generation of hierarchical feature maps at different stages [48, 120]. As illustrated in Figure 9, the encoder takes a medical input volume $\mathcal{X} \in \mathcal{R}^{H \times W \times D \times S}$, where H, W, and D represent the spatial height, width, and depth, respectively, and C is the number of channels. A 3D token with a patch resolution of (H', W', D') has a dimension of $H' \times W' \times D' \times S$. The patch partitioning layer creates a sequence of 3D tokens with size $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$ that are projected into a C-dimensional space via an embedding layer. To efficiently model token interactions, we partition the input volumes into non-overlapping windows and compute local self-attention within each region. Specifically, at layer l, we use a window of size $M \times M \times M$ to evenly divide a 3D token into $\lceil \frac{H'}{M} \rceil \times \lceil \frac{W'}{M} \rceil \times \lceil \frac{D'}{M} \rceil$ windows. The encoder block outputs in layers l and $l + 1$ are computed as shown in Figure 9c, where W-MSA and SW-MSA denote regular and window partitioning multi-head self-attention modules, respectively. A 3D cyclic-shifting is also adopted for efficient batch computation of shifted windowing [85, 120].

3.2.3 CNN Branch for 3D Inputs

Our CNN encoder branch is composed of a series of convolutional layers with a skip connection to improve network stability. The use of convolutional layers in the encoder helps to detect local patterns and features such as edges and corners in the image. Specifically, it first applies a convolutional layer with 36 $(1 \times 1 \times 1)$ spatial filters with stride 1 to the input data, and then passes it through four down-sampling residual blocks. Each residual block consists of one tri-linearly down-sampled layer followed by two 3D convolutional layers. The first convolutional layer has a $1 \times 1 \times 1$ spatial filter with stride 1 in each direction while the second convolutional layer uses

$3 \times 3 \times 3$ filters with the same stride. A skip connection used in ResNet [51] is applied between the outputs of the first and second convolutional blocks.

3.2.4 Cross-Fusion for Two Branches

To fully utilize both local and global features in our encoder, we use a parallel structure with a CNN and transformer blocks at each stage. To fuse these features, we introduce a cross-fusion module (shown in Figure 9b). This module takes two inputs with the same shape, $F_i \times H_i \times W_i \times D_i$, for the i-th stage, where F_i is the channel size. The module concatenates these two inputs and passes them through two layers of $3 \times 3 \times 3$ convolution with residual connections. The output of this module is a fused feature map with the same shape as the input, which is then used as input for the proceeding decoding operations. This simple and efficient module allows us to combine the strengths of both CNN and transformer blocks in our encoder.

3.2.5 Attention-enabled Decoder

Standard CNN architectures gradually down-sample the feature-map grid to capture a large receptive field and semantic contextual information. However, reducing false-positive predictions for small, variably shaped objects remains challenging. To address this issue, existing segmentation frameworks rely on separate object localization models. Here, we propose integrating AGs into a standard CNN model [63] to achieve the same objective without training multiple models or adding extra parameters. Unlike localization models in multi-stage CNNs, AGs progressively suppress feature responses in irrelevant background regions without the need to crop regions of interest between networks.

Additive AGs are employed to modulate feature responses through skip connections, to determine a gating vector for each pixel enabling focus on relevant regions at each multi-scale level. Although more computationally intensive than multiplicative attention, previous studies [101] have shown that additive AGs achieve superior predictive accuracy. An additive vector concatenation-based attention was adapted, in which the output of the n^{th} multi-scale encoding convolutional block (x^l) was added to the output of the $(n+1)^{th}$ multiscale decoding convolutional block (x^g), and the *ReLu* activation function applied to the combined activations. The input undergoes a channel-wise $1 \times 1 \times 1$ convolutional layer, batch normalization layer, and sigmoidal activation layer is then multiplied and concatenated to the input of the n^{th} multi-scale level decoding convolutional block. Figure 9d illustrates the attention gating mechanism.

3.3 Data Acquisition and Preprocessing

All experiments were implemented on a server equipped with 8 Nvidia A100 GPUs, each with 40 GB of memory. All experiments were conducted in the PyTorch framework in Python 3.8.13, and each model was trained on a single GPU. Data augmentation was applied during training.

3.3.1 Institutional dataset: Pelvic Multi-Organ Segmentation Dataset

Planning CT and structure datasets for 300 prostate cancer patients were retrospectively selected. The 300 cases were randomly split into a training set of 225 cases, a validation set of 30 cases, and a testing set of 45 cases. The testing dataset was "held out" and therefore “unseen" relative to CT scans used for training and validation. All CT scans were resampled into a fixed voxel spacing of $[1.0 \times 1.0 \times 1.5]$ mm^3 [81], and a Hounsfield unit (HU) range of $[-50, 150]$ was used and normalized to $[0, 1]$. Subsequently, each CT scan was cropped to a $192 \times 192 \times 64$ voxel patch around the prostate/bladder/rectum regions, used in both training and inference for the 3D models. The models were trained for 200 epochs using the AdamW (Adaptive Moment Estimation Weighted, a variant of Adam where the weight decay is performed only after controlling the parameter-wise step size) optimizer with an initial learning rate of $5e^{-4}$. An exponential learning rate scheduler with a warmup of 5 epochs was applied to the optimizer. Random flip, rotation, and intensity scaling were used as augmentation transforms, with probabilities of 0.1, 0.1, and 0.2, respectively. The training datasets were increased by a factor of approximately 175 using data augmentation. The training process for 200 epochs required approximately 16.5 *h*.

Ground-truth segments were available for all image datasets consisting of physician-drawn contours for the prostate gland (target) and surrounding normal tissues (bladder and rectum). The automatic contours generated by our network were compared to those of the ground-truth contours to evaluate the performance of the network.

3.3.2 Public Dataset: CT Organ Segmentation Dataset (CT-ORG)

A publicly available dataset (CT-ORG) [109] was used for the training and evaluation of our network for the auto-segmentation of other organs. Details of the CT-ORG dataset are provided by Rister et al. [109] The dataset consisted of 100 CT scans, each of which included manual (ground-truth) contours of the lungs, liver, bladder, kidney, and pelvic bones. The first 19 CT cases were held out and used solely for testing. The remaining 81 cases were used for training following the process of Rister et al [109]. Each CT dataset was resampled with a voxel size of [2×2×5] mm^3, and input patches of size $128 \times 128 \times 64$ were applied. Each CT dataset was truncated to a HU range of $[-1000, 1000]$ and normalized $[-1, 1]$ over this range. The same augmentation and training strategy as the institutional dataset was applied to the CT-ORG dataset. The training process for 200 epochs required approximately 10.5 *h*.

3.4 Experiment Setup

While Chapter 1 covers general definitions of loss functions and evaluation metrics, this section will provide specific descriptions for the 3D multi-organ segmentation setting.

3.4.1 Loss Functions

We utilize cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = \frac{1}{C}\sum_{k=1}^{i} y_k \log(\hat{y}_k) \tag{3.1}$$

and Dice loss:

$$\mathcal{L}_{\text{Dice}} = \frac{1}{C}\sum_{k=1}^{C}(1 - \frac{2\sum_{i\in I} y_k^i \hat{y}_k^i}{\sum_{i\in I} y_k^i + \sum_{i\in I} \hat{y}_k^i}) \tag{3.2}$$

for training, where C is the number of classes, I represents the whole volume of a 3D medical image input, y_k and $\hat{y}_k$ are the ground truth mask and the predicted segregation from the model of class k, respectively. The overall loss function was cast as an equally weighted summation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}. \tag{3.3}$$

3.4.2 Evaluation Metrics

We use the DSC and 95^{th} percentile HD to evaluate the accuracy of segmentation in our experiments. DSC evaluates the overlap of the predicted and ground truth segmentation map in 3D:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|}, \tag{3.4}$$

where P indicates the predicted segmentation map and G denotes the ground truth. A DSC of 1 indicates a perfect segmentation while 0 indicates no overlap at all. HD measures the largest symmetrical distance between two segmentation maps:

$$\text{HD} = \max\{\sup_{p\in P}\inf_{g\in G} d(p,g), \sup_{g\in G}\inf_{p\in P} d(p,g)\}, \tag{3.5}$$

where $d(\cdot)$ represents the Euclidean distance, sup and inf denote supremum and infimum, respectively.

We also include ASD, an average of all the distances from points on the surface of the predicted segmentation mask to the surface of the ground truth mask:

$$\text{ASD} = \frac{1}{|S(P)|}\sum_{p\in P} d(S(p), S(G)), \tag{3.6}$$

where $d(S(p), S(G))$ is the shortest distance of a predicted voxel $S(p)$ to the set of ground truth surface voxels, $S(G)$.

3.4.3 Methods for Comparison

The performance of SwinAttUNet was compared against multiple state-of-the-art segmentation models. For FCN-based models, V-Net [95], ResUNet [41], AttUNet [101], and nnUNet [55] are used for comparison in both institutional and public dataset. For Transformer-based models, UNETR [48] and SwinUNETR [120] were selected for comparison. P values were computed using the Mann–Whitney U-test [93] to evaluate statistical significance between contours predicted using SwinAttUNet and the next highest performing network. The significance level was set at 0.05, where $p < 0.05$ indicates a statistically significant difference between the two networks.

To validate the effectiveness of the SwinAttUNet architecture, we ran an ablation experiment using the institutional pelvic dataset. We first removed the AG to demonstrate its benefits in the decoding process. We then replaced the parallel encoder with only the CB or TB and compared each iteration against the full SwinAttUNet network.

Oragn	Method	DSC (%) ↑	HD (mm) ↓	ASD (mm) ↓
Prostate	w/o AG	86.12 ± 1.45	5.23 ± 1.50	1.51 ± 0.63
	w/o CB	85.36 ± 2.43	6.15 ± 1.46	1.62 ± 0.67
	w/o TB	84.69 ± 2.51	5.76 ± 1.43	1.56 ± 0.59
	Full model	**86.54** ± **1.21**	**5.06** ± **1.42**	**1.45** ± **0.57**
Bladder	w/o AG	93.72 ± 4.31	3.18 ± 1.26	0.85 ± 0.61
	w/o CB	93.51 ± 3.32	3.25 ± 1.33	0.93 ± 0.43
	w/o TB	93.24 ± 4.16	3.42 ± 1.37	0.86 ± 0.36
	Full model	**94.15** ± **1.17**	**3.16** ± **0.93**	**0.82** ± **0.12**
Rectum	w/o AG	86.31 ± 2.12	5.74 ± 1.94	1.53 ± 0.51
	w/o CB	86.25 ± 1.83	6.11 ± 2.07	1.61 ± 0.68
	w/o TB	85.49 ± 2.08	5.83 ± 1.85	1.53 ± 0.51
	Full model	**87.15** ± **1.68**	**5.54** ± **1.63**	**1.42** ± **0.38**

Table 5: Ablation Study: DSC, HD, and ASD with the different settings for SwinAttUNet on the institutional pelvic dataset. Shown are mean ± SD for three runs for each setting for the prostate gland, bladder, and rectum. The most accurate results are shown in bold font. Abbreviations: AG, Attention Gate; CB, CNN Branch; TB, Transformer Branch.

3.5 Results and Discussion

3.5.1 Overview of Qualitative Comparisons

Qualitative comparisons of auto-contours generated with SwinAttUNet, and other networks are presented in Figures 10 and 11. Figure 10 shows results for four example cases based on the institutional pelvic dataset. Contours are shown in the axial view for the ground truth (GT, Row 2), SwinAttUNet (Row 3), and competing networks (Rows 4-9), for the prostate gland (blue), bladder (green), and rectum (red). Careful inspection of the shapes of these contours and the boundary distances between the different organs (relative to the ground truth segments, Row 2) shows that the SwinAttUNet performs better than all other networks for all four example cases. Figure 11 shows results for five example cases based on the public dataset (CT-ORG). Contours are depicted

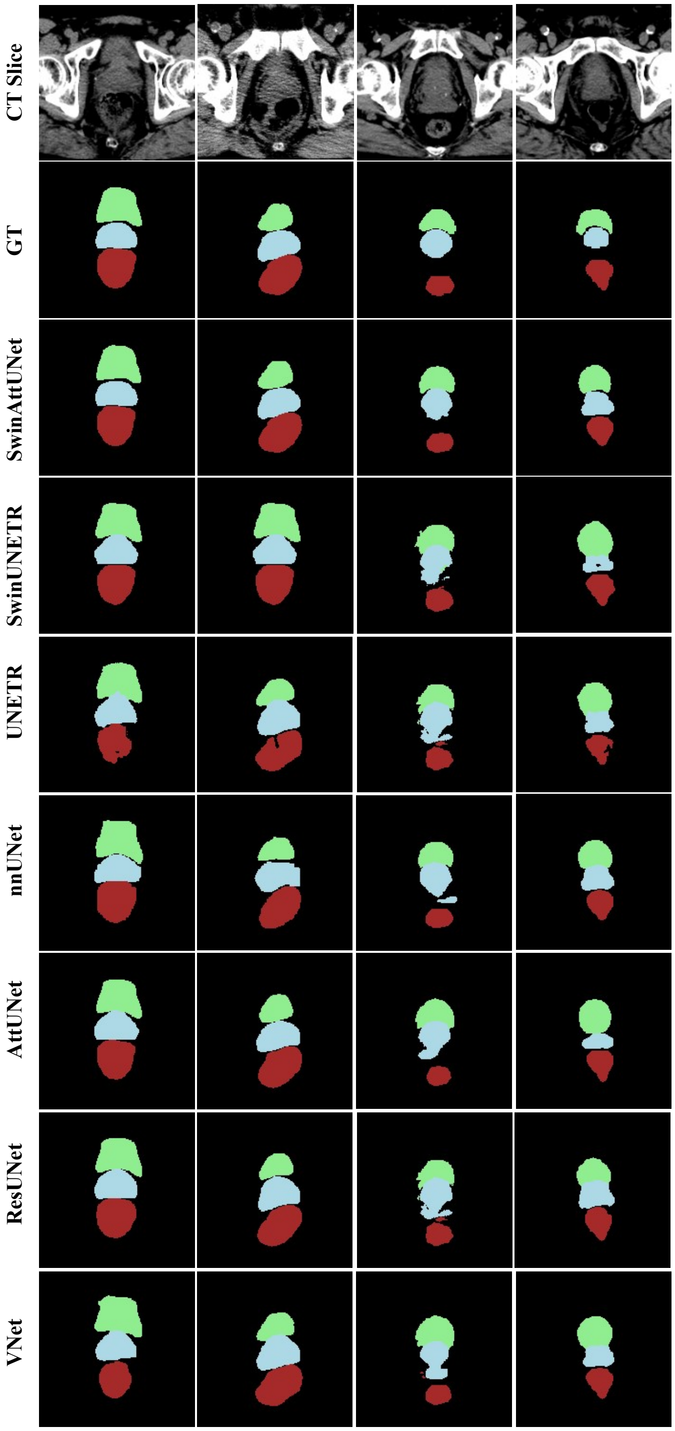

Figure 10: Segmentation result from the institutional pelvic dataset. The input CT image of a central slice (Row 1), the ground truth (Row 2), and the predicted segmentation from the SwinAttUNet (Row 3), and all competing networks (Rows 4-9): prostate (blue), bladder (green), and rectum (red).

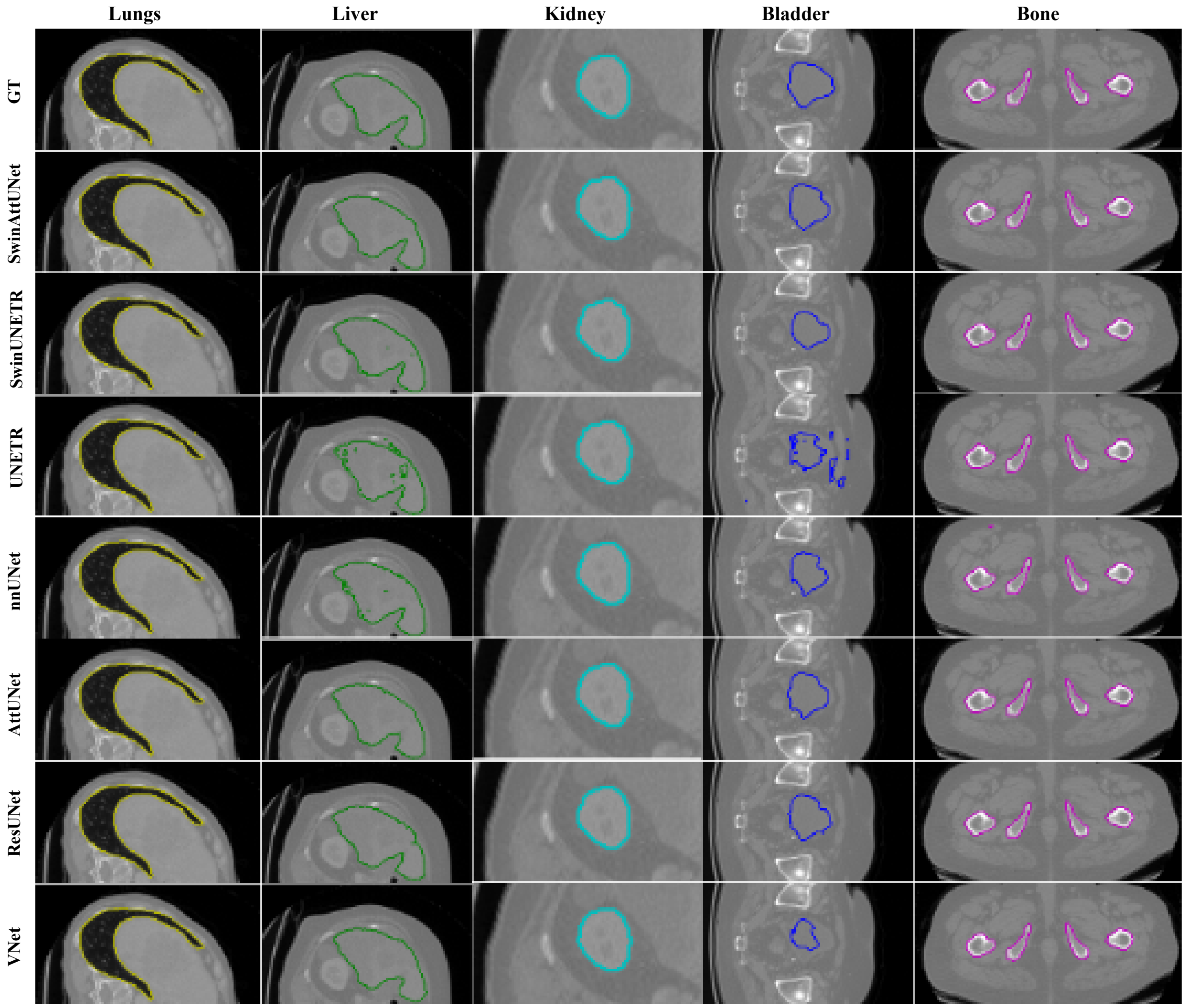

Figure 11: Segmentation result from CT-ORG dataset. The selected region of interest of each organ from manual contours (Row 1), the SwinAttUNet (Row 2), and all competing networks (Rows 3-8): lungs (yellow), liver (green), kidney (cyan), bladder (blue), and bones (purple).

Oragn	Method	DSC (%) ↑	HD (mm) ↓	ASD (mm) ↓
Prostate	V-Net	83.27 ± 2.71	7.75 ± 2.58	2.12 ± 0.73
	ResUNet	84.15 ± 2.61	5.79 ± 1.63	1.74 ± 0.89
	AttUNet	84.26 ± 2.54	5.81 ± 1.56	1.58 ± 0.65
	nnUNet	84.12 ± 2.68	5.83 ± 2.01	1.81 ± 0.79
	UNETR	82.51 ± 4.46	8.92 ± 2.65	2.34 ± 1.01
	SwinUNETR	85.71 ± 2.32	6.10 ± 1.42	**1.40** ± **0.65**
	SwinAttUNet (ours)	**86.54** ± **1.21**	**5.06** ± **1.42**	1.45 ± 0.57
P-values		< 0.001	< 0.001	0.076
Bladder	V-Net	91.56 ± 5.21	6.75 ± 2.01	1.62 ± 0.52
	ResUNet	92.65 ± 4.52	4.46 ± 1.84	1.13 ± 0.24
	AttUNet	93.31 ± 4.23	3.25 ± 1.21	0.87 ± 0.54
	nnUNet	93.46 ± 5.03	4.83 ± 1.59	1.16 ± 0.46
	UNETR	89.37 ± 5.67	6.34 ± 2.56	1.78 ± 0.67
	SwinUNETR	93.62 ± 3.25	3.22 ± 1.14	0.91 ± 0.34
	SwinAttUNet (ours)	**94.15** ± **1.17**	**3.16** ± **0.93**	**0.82** ± **0.12**
P-values		< 0.001	< 0.001	< 0.001
Rectum	V-Net	83.71 ± 3.52	7.12 ± 2.54	2.11 ± 0.61
	ResUNet	86.02 ± 2.34	6.31 ± 2.24	1.58 ± 0.46
	AttUNet	86.63 ± 2.01	5.81 ± 1.95	1.44 ± 0.49
	nnUNet	86.53 ± 2.18	6.14 ± 2.35	1.61 ± 0.62
	UNETR	82.16 ± 4.87	9.76 ± 2.45	2.43 ± 1.10
	SwinUNETR	85.52 ± 2.24	6.12 ± 1.97	1.47 ± 0.61
	SwinAttUNet (ours)	**87.15** ± **1.68**	**5.54** ± **1.63**	**1.42** ± **0.38**
P-values		< 0.001	< 0.001	< 0.001

Table 6: Quantitative performance comparison on the institutional pelvic dataset in terms of DSC, HD, and ASD metrics. The values represent the mean performance (and standard deviation) of 3 runs for each setting. The Mann–Whitney U-Test statistical analysis is presented to compare the SwinAttUNet with the other networks. The best results are bolded.

Oragn	Method	DSC (%) ↑	HD (mm) ↓	ASD (mm) ↓
Liver	V-Net	94.13 ± 2.54	6.48 ± 2.32	1.73 ± 0.63
	ResUNet	94.81 ± 1.81	5.74 ± 3.43	1.42 ± 0.87
	AttUNet	95.23 ± 1.72	4.53 ± 1.67	1.26 ± 0.75
	nnUNet	94.78 ± 1.95	6.21 ± 4.02	1.38 ± 0.84
	UNETR	94.01 ± 2.34	6.83 ± 4.21	5.58 ± 3.54
	SwinUNETR	94.81 ± 2.34	4.85 ± 2.63	1.97 ± 1.65
	SwinAttUNet (ours)	**96.16 ± 0.76**	**2.73 ± 1.19**	**1.08 ± 0.24**
P-values		< 0.001	< 0.001	0.004
Bladder	V-Net	83.24 ± 11.75	7.68 ± 4.32	2.56 ± 0.97
	ResUNet	82.48 ± 12.24	9.73 ± 6.85	3.12 ± 1.54
	AttUNet	84.87 ± 11.86	8.56 ± 6.53	2.15 ± 1.17
	nnUNet	85.26 ± 12.58	10.21 ± 8.57	2.53 ± 2.64
	UNETR	82.13 ± 13.65	10.02 ± 5.84	2.86 ± 2.13
	SwinUNETR	83.67 ± 13.15	8.76 ± 6.21	2.24 ± 1.35
	SwinAttUNet (ours)	**88.62 ± 7.91**	**8.23 ± 8.01**	**1.78 ± 1.21**
P-values		< 0.001	0.673	< 0.001
Lungs	V-Net	95.63 ± 4.36	15.61 ± 8.84	2.89 ± 0.86
	ResUNet	95.82 ± 5.27	6.64 ± 15.76	3.12 ± 4.42
	AttUNet	96.87 ± 5.13	5.99 ± 11.97	3.31 ± 2.37
	nnUNet	95.63 ± 6.85	8.57 ± 5.12	3.53 ± 6.56
	UNETR	93.68 ± 13.64	15.25 ± 19.40	8.37 ± 9.06
	SwinUNETR	95.99 ± 9.30	**4.95 ± 4.37**	2.56 ± 1.63
	SwinAttUNet (ours)	**97.90 ± 0.80**	5.13 ± 4.11	**1.88 ± 1.45**
P-values		< 0.001	0.893	< 0.001
Kidney	V-Net	88.15 ± 3.25	4.45 ± 1.87	2.14 ± 3.32
	ResUNet	92.03 ± 3.40	3.26 ± 1.32	0.95 ± 0.75
	AttUNet	92.85 ± 3.73	**2.12 ± 1.14**	0.83 ± 0.58
	nnUNet	93.41 ± 3.92	3.11 ± 5.73	1.01 ± 0.78
	UNETR	88.96 ± 4.88	6.36 ± 9.03	3.18 ± 3.59
	SwinUNETR	91.87 ± 3.41	3.37 ± 1.24	0.94 ± 0.66
	SwinAttUNet (ours)	**93.74 ± 2.25**	2.29 ± 1.47	**0.71 ± 0.43**
P-values		0.003	0.653	< 0.001
Bone	V-Net	86.45 ± 2.17	8.76 ± 3.21	2.22 ± 3.28
	ResUNet	88.61 ± 4.95	5.58 ± 5.87	1.44 ± 1.23
	nnUNet	88.63 ± 4.57	5.67 ± 6.12	2.19 ± 2.67
	UNETR	86.85 ± 6.39	8.78 ± 9.03	4.72 ± 4.90
	SwinUNETR	88.97 ± 4.80	5.63 ± 6.03	2.43 ± 2.44
	SwinAttUNet (ours)	**89.31 ± 3.87**	**5.31 ± 1.25**	**1.21 ± 1.11**
P-values		< 0.001	0.005	0.023

Table 7: Quantitative analysis for CT-ORG dataset: the table shows statistics for the DSC, HD, and ASD for the proposed SwinAttUNet, VNet, ResUNet, AttUNet, nnUNet, UNETR, and SwinUNETR. The statistical analysis of the Mann–Whitney U-test is also efficiently presented to compare the SwinAttUNet with each competing network. The best performance is bolded.

for the ground truth (GT, Row 1), SwinAttUNet (Row 2), and competing networks (Rows 3-8) for the lungs (yellow), liver (green), kidney (cyan), bladder (blue), and pelvic bones (purple). While all networks produce accurate contours of the liver, lung, kidney, and bones, the SwinAttUNet is shown to produce the best contours for all organs, including the bladder where discrepancies were noted with the other networks relative to the ground-truth segments.

3.5.2 Ablation Study for Different Modules of SwinAttUNet

To assess the contribution of the AG, CB, and TB on the segmentation performance, a comparison was performed between the results obtained with the SwinAttUNet (full model) and the network configurations without AG, CB, or TB. Table 5 presents the segmentation results for these three different experiments. The SwinAttUNet (full model) shows superior results for all metrics, DSC, HD, and ASD for the prostate, bladder, and rectum. The contribution of the various modules of the SwinAttUNet is demonstrated by inferior results when the AG, CB, or TB are removed from the network architecture, justifying the need for each module toward the overall accuracy of the SwinAttUNet network.

3.5.3 SwinAttUNet on Institutional Dataset for Pelvic Organ Segmentation

Quantitative results for the SwinAttUNet and other networks trained on the institutional dataset for the segmentation of pelvic organs are provided in Table 6. Data are shown for the DSC (%), HD (*mm*), and ASD (*mm*) for the prostate, bladder, and rectum. *p*-values are also included for comparison between the SwinAttUNet and the next highest accuracy network at the 0.05 significance level. For the DSC comparison, the SwinAttUNet outperforms all other networks with values of 86.5% (prostate), 94.2% (bladder), and 87.2% (rectum). The HD95 (mm) values were also the lowest for our SwinAttUNet relative to other networks. Statistically significant differences ($p < 0.001$) were observed in the DSC and HD values for our network (SwinAttUNet) versus SwinUNETR for all organs. Apart from the prostate, SwinAttUNet ASD (mm) values outperformed those of all other networks.

For the prostate, the ASD values were 1.40 *mm* (SwinUNETR) and 1.45 mm (our SwinAttUNet), however, the difference was not statistically significant (p = 0.076). Moreover, the standard deviation of the prostate ASD with SwinAttUNet (0.57 *mm*) was lower than that of SwinUNETR (0.65 *mm*).

3.5.4 SwinAttUNet on the Public CT-ORG Dataset for Multi-organ Segmentation

Quantitative results for the SwinAttUNet and other networks trained on the CT-ORG dataset for the segmentation of multiple organs are provided in Table 7. Data are shown for the DSC (%), HD (*mm*), and ASD (*mm*) for the lungs, liver, kidneys, bladder, and pelvic bones. DSC values are consistently the highest for the SwinAttUNet versus all other networks with values of 97.9% (lungs), 96.2% (liver), 93.7% (kidneys), 88.6% (bladder), and 89.3% (pelvic bones). Statistically significant DSC differences ($p < 0.001$) were observed for the SwinAttUNet relative to the SwinUNETR network. Moreover, DSC standard deviations were significantly reduced on segments produced with SwinAttUNet relative to other networks. For instance, the bladder DSC SD was 7.9 mm for SwinAttUNet, while it was > 11.5 *mm* for all other networks. HD (*mm*) values were the lowest for our SwinAttUNet relative to other networks for the liver and pelvic bones with statistical significance achieved (against SwinUNETR) for the liver ($p < 0.001$) and pelvic bones (p = 0.005). For the lungs, SwinAttUNet HD mean value was 5.13 *mm* while it was slightly better with SwinUNETR (4.95 *mm*) though the difference was not statistically significant (p = 0.89). For the kidneys, SwinAttUNet HD mean value was 2.29 *mm* while it was 2.12 *mm* for the AttUNet network but not statistically different (p = 0.65). For the bladder, the HD mean value was 7.68 *mm* for the V-Net slightly better than 8.23 *mm* for the SwinAttUNet network but not statistically different

(p = 0.67). ASD values were lowest for all organs with our SwinAttUNet network with statistical significance consistently achieved. ASD SDs were also significantly improved with SwinAttUNet. For instance, ASD SDs for the liver were reduced to 0.24 *mm* with SwinAttUNet compared with all other networks where the SDs were generally > 0.7 *mm*, suggesting lower variability and higher consistency in the predicted contours with our network.

3.5.5 Discussion

In this work, we propose a U-shaped hierarchically fusing architecture called SwinAttUNet for 3D CT-based multi-organ segmentation. The SwinAttUNet consists of three main components: a convolutional encoder branch for extracting fine local features at different resolutions, a Swin Transformer encoder branch in parallel for enriching global information at each resolution level, and a set of AG-regulated, up-sampling convolutional blocks for reconstruction of features into an N-class segmentation. Our network is novel in that the transformer layers effectively capture global information in parallel with the CNN layers for each resolution level, overcoming the receptive field limitations of pure fully convolutional networks (FCNs). Additionally, the novel AGs enable effective interaction of extracted features from different resolution levels, as evidenced by the ablation study. The proposed network demonstrates promising segmentation performance compared to current state-of-the-art methods for auto-segmentation of organ contours in multiple regions of the body including the pelvis, thorax, and gastrointestinal. The superior DSC, HD, and ASD results of our proposed network highlight the advantages of parallelizing the CNN and Swin Transformer layers in the encoding stage for CT-based multi-organ segmentation.

Relative to our network other CNN-based models with more complex architectures and ground truth segmentations based on multi-modal imaging information were able to achieve similar accuracies. An example of this is a study by Dong et al. [67], where the investigators utilized a

Cycle-GAN for 3D CT-to-synthetic MRI synthesis and trained the segmentation network on the synthetic MRI scans. They reported Dice scores of 0.87 ± 0.04 for the prostate, and 0.95 ± 0.03 for the bladder using 140 pelvic image datasets. Similarly, other investigators used networks such as GAN for CT-to-sMRI synthesis [102] (with Dice scores of 0.90 ± 0.05 for the rectum) and 2D organ localization networks [5] (with Dice scores of 0.95 ± 0.02 for bladder). In the context of postoperative prostate cancer, Balagopal et al. [7] developed a deep learning network (2D U-Net) for auto-segmentation of the clinical target volume (CTV) incorporating uncertainty. The training dataset consisted of 340 patients with post-operative prostate cancer, with ground-truth contours drawn by physicians. A DSC value of 0.87 was reported for a holdout dataset (50 patient CT images). Balagopal et al.[6] also developed a deep-learning network (based on a 3D-CNN), PSA-Net, for segmentation of the CTV trained to incorporate differences in physician preferences during segmentation. For training, 373 postoperative prostate cancer CT image datasets were employed. Questions such as consistency in physician contouring preferences and whether inter-user variation in segmentation affects treatment outcomes were addressed. DSC values of 0.87 were reported for their network.

There are a few limitations to be noted. We trained our network on two independent (institutional and public datasets) training datasets because the ground-truth labels/contours were not available for the same organs on these datasets. If contours were available for the same set of organs, it would have become feasible to train the network with just one training dataset, which may be more practical for clinical application. The generalization error of a network is best tested using "unseen" datasets from an independent institution, as it tests the robustness of the network to variation associated with multiple factors, such as image intensity and contrast, patient anatomy, inter-observer differences in ground truth contours of expert annotators, etc.

As part of future research, we intend to evaluate the network using unseen datasets from independent institutions. We will also incorporate advanced techniques/networks to enhance the segmentation accuracy of SwinAttUNet. For instance, we propose to extend the parallel CNN and Transformer into the decoding process, which has the potential to increase segmentation accuracy. Our network is efficient. Apart from the training phase (which requires $> 10\ h$ but is done before clinical application), the network is fast, requiring about 5 s/case for routine multi-organ contour generation, thereby facilitating auto-segmentation for procedures such as on-table adaptive treatment. We are investigating techniques to automatically detect and correct outliers from either manually (user-defined) or automatically generated contours using this network. These tools are likely to be of value toward the overall quality assurance of target and normal organ segmentation in radiation treatment planning.

3.6 Conclusion

SwinAttUNet, an advanced deep learning architecture, combines convolutional and transformer-based methods for the automated segmentation of organs in pelvic, thoracic, and gastrointestinal regions. Integrating a shifted-window (Swin) transformer with a convolutional U-Net, it features a parallel encoder, cross-fusion block, and attention-enhanced CNN decoder. Trained on diverse CT datasets, SwinAttUNet excels in accuracy, surpassing existing models with its adept handling of both local and global anatomical features. This proficiency positions it as an efficient tool for critical applications like radiation treatment planning, showcasing its potential in multi-organ, 3D-CT auto-segmentation.

While SwinAttUNet demonstrates exceptional capabilities in 3D multi-organ medical image segmentation, its training on a limited number of medical image data in only one modality (CT) presents a notable limitation. Building upon this, the emerging need for robust segmentation across diverse medical imaging modalities presents a new challenge.

In the diverse landscape of medical imaging, where input ranges from multiple types of Magnetic Resonance (MR) imaging to CT scans, automatic segmentation algorithms face the challenge of maintaining consistent performance across different modalities due to the conventional requirement for spatially aligned and paired images. We introduce the Multi-Modal Segmentation (MulModSeg) strategy to address this challenge in the following Chapter 4, specifically designed for unpaired CT/MR images. It incorporates two key innovations: a modality-specific text embedding via CLIP model that adds modality awareness to existing segmentation frameworks without significant computational overhead or structural modifications, and an alternating training method that facilitates the integration of essential features from unpaired images.

CHAPTER 4 MULMODSEG: ENHANCING UNPAIRED MULTI-MODAL SEGMENTATION FOR MEDICAL IMAGES

4.1 Introduction

4.1.1 Background Significance

Medical image segmentation leverages multiple imaging techniques, such as Computed Tomography (CT) and Magnetic Resonance (MR) Imaging, to provide comprehensive views of tissues or organs for disease diagnosis and surgical planning [73]. Different modalities offer unique advantages; MR provides superior soft tissue contrast, while CT delivers better bone detail and higher spatial resolution [94]. Recent advancements in convolutional [110, 91] and transformer-based neural networks [48, 47, 71] have achieved competitive segmentation precision. However, while humans can easily identify features across modalities, algorithms trained on single modalities struggle with segmenting multiple modalities. This leads to performance inconsistencies when tested on different image types due to data variability, stemming from factors such as varying imaging methods, scanners, acquisition settings, or patient conditions [159]. Training separate models for each modality would be straightforward, but it would require a massive amount of annotated data and could fail to leverage inter-domain information.

4.1.2 Related Work

To address this issue, researchers proposed several multi-modal medical image segmentation methods [161, 149, 92, 73, 162]. The first group of methods aims to produce better segmentation by simultaneously utilizing information from multiple modalities. In this context, techniques such as input/layer/decision-level fusion [161], modality-specific representation [160], and hyperdense connections [29] have been employed to enhance segmentation by integrating information from diverse sources more effectively than single-modality methods. Typically, different modalities are treated as separate inputs for the model, which then generates combined inputs or learns a common representation. However, these methods often require spatially aligned, paired images from the same patient, a condition rarely met due to misalignments and variations in unpaired images, thus compromising performance [32].

The exploration of multi-modal learning for medical image segmentation from unpaired CT and MR scans has catalyzed significant innovations by leveraging the unique attributes of each modality without depending on paired datasets. Research by Dou et al. [32] and Jiang et al. [59] has led to the development of compact, efficient architectures that share convolutional kernels between modalities and incorporate modality-specific normalization alongside innovative loss functions inspired by knowledge distillation, enhancing segmentation accuracy across diverse data types. These advancements underscore the potential of dual-stream architectures [37], attention mechanisms [139], and adversarial training [98] in improving segmentation performance. Yet, challenges persist in harnessing shared cross-modality information due to additional preprocessing, which impedes learning across significant domain shifts, such as those between MR and CT images. Other methods like multiple feature extractors per modality, suggested by the X-shaped architecture [124], add overhead and necessitate broader clinical adaptations. The shift towards synthetic image generation [17] and semi-supervised learning [82] highlights a growing reliance on unlabeled data to address the paired image scarcity, with methods like CycleGAN [165] showing promise in synthesizing cardiac MR images from CT scans. However, these approaches necessitate significant modifications to existing segmentation frameworks or the introduction of additional complex steps for self-supervised training or synthesis, making their application challenging.

4.1.3 Our Contribution

In this study, we introduce a versatile Multi-Modal Segmentation (MulModSeg) strategy, designed for seamless integration of modality-conditioned text embedding with any encoder-decoder architecture. This approach aims to enhance multi-modality medical image segmentation without major architectural modifications during supervised training. Our framework features two key innovations: a modality-specific text embedding via the frozen CLIP [24] text encoder that introduces modality awareness to existing segmentation frameworks (as shown in Figure 12), and an alternating training algorithm that facilitates the integration of essential features from unpaired images. The MulModSeg strategy self-adjusts its decoder embedding layers to generate precise segmentation outputs, inspired by the CLIP-driven universal model concept [83]. Moving beyond its initial application within the CT modality for multi-class segmentation [83], this study expands the use of modality-conditioned text embeddings for tasks across multiple modalities, specifically CT and MR. To enable straightforward, one-pass end-to-end supervised training, we propose an Alternating Training (ALT) strategy (see Alg. 1), effectively managing batched CT and MR samples sequentially. These two designs allow the use of a single encoder-decoder structure to accurately segment images across both modalities. Moreover, this modality-conditioned text embedding and alternating training method can be easily integrated into popular FCN-based and Transformer-based medical image segmentation networks, such as UNet [110] and SwinUNETR [47].

Our contributions are three-fold: (1) We propose the MulModSeg method using modality-conditioned text embedding. This adds modality awareness to existing encoder-decoder segmentation frameworks via a frozen CLIP text encoder. It does so without requiring major architectural modifications or significant computational overhead. (2) We introduce an alternating training procedure to integrate essential features from unpaired CT/MR images. This enables straightforward, one-pass, end-to-end supervised training across both CT and MR modalities. It streamlines the training process and improves segmentation performance. (3) We demonstrate the superior performance of MulModSeg over previous methods by conducting extensive performance evaluations on abdominal multi-organ segmentation using AMOS [57] dataset and cardiac substructure segmentation using MMWHS [166] dataset.

4.2 More Related Work

4.2.1 Multi-Modality Learning in Medical Imaging

In recent years, several deep learning architectures have been proposed for image segmentation, achieving remarkable performance [95, 20, 47, 48]. Among them, UNet [110] stands out as the most popular and is often used as a baseline for developing better-performing models. More recently, variations based on vision transformers (ViT) [31] and Swin transformers [86], such as TransUNet [20], UNETR [48], and SwinUNETR [47], have shown superior performance compared to previous versions of UNet. In the clinical imaging field, multi-modality learning presents a similar, yet challenging, task. Multiple medical image domains have been leveraged for synthetic image generation using Generative Adversarial Networks (GAN) [27, 145] or for multi-modal image segmentation [45, 133]. Notably, Zhang et al. [150] merged these tasks in a cross-task feedback fusion GAN that first generates synthetic CT images and then performs multi-modal segmentation, using cross-domain information to enhance performance but requires registered paired images. Cross-modality segmentation approaches have gained traction in overcoming the limitation of needing aligned medical images and exploiting inter-domain features during training. Zheng et al. [155] were pioneers in this area, using shape priors learned from an assistant modality to improve segmentation on a target modality through marginal space learning. Valindria et al. [125] developed dual-stream encoder-decoder models with separate branches for each modality, implementing weight-sharing techniques to extract cross-modality features. Additionally, some researchers have experimented with normalization layers to enhance model generalization [115, 163]. For example, Pan et al. [103] introduced the IBN-Net, which leverages both Instance and Batch Normalization to capture appearance changes and content information. Segu et al. [115] proposed training adhoc Batch Normalization layers to collect domain-dependent statistics, mapping modalities onto a shared latent space. Advancements in synthetic image generation have led to several works that assist segmentation models with prior image translation [22, 74, 152, 163]. Zhang et al. [152] sought to improve segmentation for modalities with limited training samples by using a GAN to reduce the appearance gap between modalities. Similarly, Li et al. [74] introduced an Image Alignment Module to minimize the appearance gap between assistant and target modalities and implemented a Mutual Knowledge Distillation scheme [53] to utilize shared knowledge across modalities. More recently, Zhou et al. [163] proposed simulating possible appearance changes in target domains through non-linear transformations to augment source-similar and source-dissimilar images.

Our framework is different from existing methods as it aims to improve segmentation accuracy by extracting modality-specific features through modality-conditioned text embedding with frozen text encoders. This reduces the need for significant modification of popular segmentation frameworks. Additionally, our technique allows for easily alternating the input modality during training, which is not possible with previous methods that predominantly rely on image translation or prior training.

4.2.2 Text Assisted Medical Image Segmentation

Text-assisted medical image segmentation has emerged as a promising approach to enhance the accuracy and efficiency of medical imaging tasks by integrating textual information with visual data. The TGANet study [122] introduced an innovative method that leverages text-based embeddings to guide segmentation models in colonoscopy procedures. By incorporating size-related and polyp number-related features in the form of text attention during training, TGANet can adapt to varying polyp sizes and numbers, thereby improving the segmentation performance compared to traditional image-only methods. Similarly, Zhong et al. [156] presented the benefits of language-driven segmentation, showing significant improvements in Dice scores and reduced training data requirements. These advancements underscore the potential of multimodal approaches in overcoming the limitations of uni-modal systems that rely solely on images. To further extend the capabilities of text-assisted segmentation, Liu et al. [80] demonstrated how frozen language models can stabilize training and enhance the latent space representation for medical vision-language models. This approach, which integrates clinical text with imaging data, has shown superior performance across various tasks, including segmentation while reducing the computational requirements. Moreover, Chen et al. [21] proposed a framework expanding the application of vision-language pretraining to 3D medical images by generating synthetic text from images using large language models. This innovative method addresses the scarcity of paired textual descriptions in the medical domain and proves effective across multiple imaging modalities. These advancements collectively illustrate the significant strides made in text-assisted medical image segmentation, highlighting its potential to improve diagnostic accuracy and efficiency in clinical settings. Despite the enhancement of the segmentation accuracy, these methods only targeted specific tasks in one

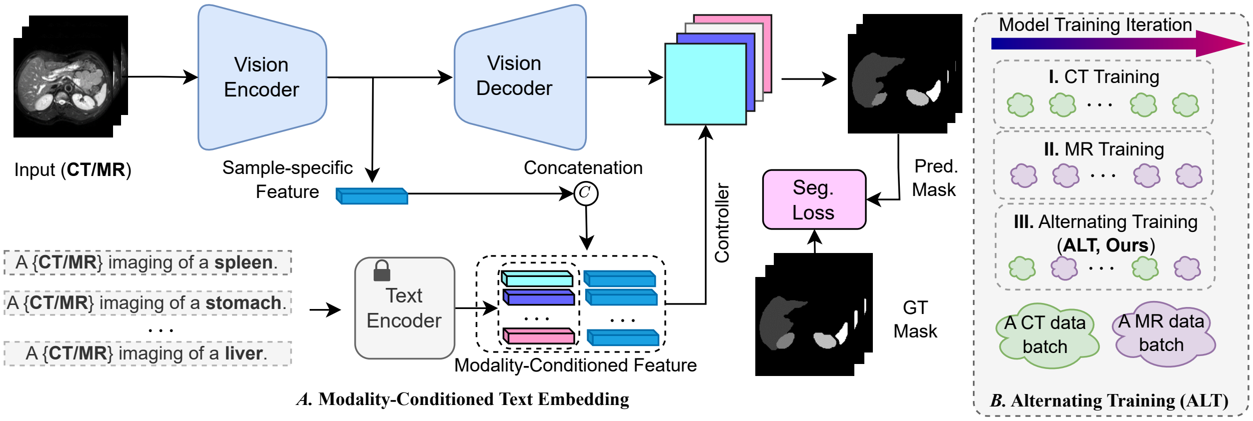
Vision Encoder
Vision Decoder
Input (CT/MR)
Sample-specific Feature
Concatenation
C
Controller
Pred. Mask
Seg. Loss
GT Mask
A {CT/MR} imaging of a spleen.
A {CT/MR} imaging of a stomach.
...
A {CT/MR} imaging of a liver.
Text Encoder
Modality-Conditioned Feature
A. Modality-Conditioned Text Embedding
Model Training Iteration
I. CT Training
II. MR Training
III. Alternating Training (ALT, Ours)
A CT data batch
A MR data batch
B. Alternating Training (ALT)

Figure 12: Schematic representation of the MulModSeg strategy for multi-modal medical image segmentation with (A) modality-conditioned text embedding and (B) alternating training (ALT).

specific modality. In this study, we explore using a frozen text encoder to encode modality-related information to the encoder-decoder architectures for multi-modal medical image segmentation.

4.3 The MulModSeg Architecture

4.3.1 Problem Definition

Consider a set of N datasets $\{D_1^{M_1}, D_2^{M_2}, \ldots, D_N^{M_N}\}$, each corresponding to a different imaging modality $M_1, M_2, \ldots, M_N$. Each dataset $D_i^{M_i} = \{(X_{ij}, Y_{ij})\}_{j=1}^{N_i}$ consists of N_i image-label pairs from modality M_i (e.g., CT or MR), where X_{ij} represents the image and Y_{ij} its associated ground truth label. Traditionally, N separate segmentation tasks might be tackled by training individual models on each of these datasets. For example, in the context of multi-class organ segmentation, such as abdominal organs, we might encounter datasets like D_1^{CT} and D_2^{MR}. However, training separate models for each modality requires a substantial amount of annotated data and fails to leverage the complementary information across modalities. To address this issue, we propose a unified approach: **MulModSeg**, a strategy designed to perform segmentation tasks across multiple modalities using a single model.

4.3.2 Overall Architecture Design

The overall architecture for MulModSeg (see Figure 12 A) has a vision branch and a text branch. The text branch first generates modality-conditioned (CT/MR) text embedding from CLIP for each category of organ, and then the vision branch takes both CT/MR scans and text embedding to predict the segmentation mask. The alternating training (ALT) is shown in Figure 12B.

4.3.3 Modality-conditioned Text Embedding Branch

Incorporating text embeddings that condition the model on the specific modality of the input image, such as CT or MR, we propose Modality-conditioned Text Embedding. It starts with a text description related to the imaging modality with a template like "A CT/MR of a [CLS]", such as "A magnetic resonance imaging of a [CLS]", where [CLS] is a concrete class of organ name, e.g., spleen, stomach. These descriptions are transformed into modality-conditioned embeddings using a pre-trained CLIP text encoder as our default setting. The generated text embeddings are then concatenated with sample-specific features from the vision encoder and passed through a multi-layer perceptron controller. This process generates weights to adjust the vision decoder's output feature maps, culminating in the generation of the final predicted segmentation mask. This technique ensures that the model adapts its behavior and segmentation strategies to the specified modality, optimizing feature extraction and segmentation accuracy for CT or MR input images.

4.3.4 Encoder-Decoder Based Vision Branch

The MulModSeg strategy effectively integrates popular U-Net-like architectures, including UNet [110] and SwinUNETR [47], into a unified encoder-decoder framework tailored for unpaired multi-modal medical image segmentation. For instance, the 3D U-Net, characterized by its dual-pathway design, combines a contracting path for downsampling and an expansive path for upsampling, forming a "U" shape. The contracting path comprises blocks of 3D convolutional layers, ReLU activation functions, and max pooling to distill contextual information into a compact form while reducing spatial dimensions. The expansive path, conversely, utilizes 3D transposed convolutions and skip connections to merge features from the contracting path, aiding in precise localization and the recovery of spatial details lost during downsampling operation. In the expansive path, a final layer of 3D convolution with $1 \times 1 \times 1$ kernels maps the feature maps to segmentation classes. MulModSeg extends the functionality of the traditional 3D U-Net by processing both standardized and normalized CT scans through the vision encoder for feature extraction. It incorporates a global averaging block to aggregate sample-specific features at the encoder's final stage, coupled with a controller layer for accurate, modality-conditioned, class-specific output adjustment of the extracted image features. Furthermore, MulModSeg innovates by substituting the conventional single final layer convolution in the decoder with a series of three sequential convolution layers with $1 \times 1 \times 1$ kernels similar to [83], whose weights are determined in the text embedding branch, enriching the model's segmentation precision and adaptability to different medical imaging modalities.

4.3.5 The Controller as A Bridge

The controller in our MulModSeg acts as a bridge connecting the textual context and image features. First, the text encoder generates embeddings for each class using the text descriptions depicted above. These embeddings (w_k) are concatenated with global image features (g) extracted by the vision encoder from CT/MR scans, forming a combined feature vector ($w_k \oplus g$). This vector is input into a multi-layer perceptron (MLP) called the text-based controller, which generates segmentation parameters: $\theta_k = \text{MLP}(w_k \oplus g)$. The vision branch preprocesses CT/MR scans for standardization and then extracts image features (F) using the vision encoder. These features are processed by a text-driven segmentor consisting of three sequential convolutional layers with $1 \times 1 \times 1$ kernels. The first two layers have 8 channels, and the last layer has 1 channel corresponding to the predicted class ($[CLS]_k$). The segmentation prediction for each class is computed as: $P_k = \text{Sigmoid}(((F * \theta_{k1}) * \theta_{k2}) * \theta_{k3})$, where $\theta_k = \{\theta_{k1}, \theta_{k2}, \theta_{k3}\}$ and $*$ for convolution. The predicted mask $P_k \in \mathbb{R}^{1\times H\times W\times D}$ denotes the foreground of each class in one *vs.* all manner.

4.3.6 Modality Driven Alternating Training

ALT (as shown in Figure 12B) is designed to enhance model training on multi-modal medical imaging data, specifically alternating between CT and MR modalities within each training iteration. The algorithm utilizes cyclic loaders for both CT and MR datasets, ensuring uninterrupted data feeding by looping back to the start once the end of a dataset is reached. During each iteration, the algorithm processes a batch from each modality: first CT, then MR by extracting images and labels, identifying the batch's modality, and feeding this information into the model to generate predictions. This methodical alternation between CT and MR batches allows for balanced exposure to both modalities, promoting model robustness and preventing bias towards either modality, thus enhancing the model's generalization capabilities across diverse medical imaging tasks. The overall ALT algorithm is summarized as follows:

4.4 Experiments

4.4.1 Experiments Setup

We aim to validate the effectiveness of the proposed MulModSeg strategy in improving segmentation performance through a series of comprehensive experiments. To achieve this, several key research questions must be addressed. ***Q1***: How effective is text embedding in improving segmentation accuracy, and which types of embedding yield the best results? ***Q2***: Do modality-

Algorithm 1 Alternating Training (ALT) with CT and MR Modalities

```
procedure TrainEpoch($CTLoader$, $MRLoader$, $model$, $optimizer$, $lossFunc$)
    $model.train()$
    $maxIter \leftarrow \max(len(CTLoader), len(MRLoader))$
    $cycleCT \leftarrow cycle(CTLoader)$
    $cycleMR \leftarrow cycle(MRLoader)$
    for $iter \in range(maxIter)$ do
        $batchCT \leftarrow next(cycleCT)$
        $batchMR \leftarrow next(cycleMR)$
        for $batch \in [batchCT, batchMR]$ do
            $images, labels \leftarrow batch['image'], batch['label']$
            $modality \leftarrow batch['modality']$
            $logits \leftarrow model(images, modality)$
            $loss \leftarrow lossFunc(logits, labels)$
            $optimizer.zero\_grad()$
            $loss.backward()$
            $optimizer.step()$
        end for
    end for
end procedure
```

conditioned text embedding and alternating training (ALT) have a positive impact on the segmentation accuracy across different organ structures and imaging modalities (CT/MR) using various backbone architectures, specifically UNet and SwinUNETR? ***Q3***: Does MulModSeg outperform existing state-of-the-art (SOTA) methods in terms of performance? ***Q4***: What is the impact of varying the ratio of CT to MR scans on the model's performance, especially in simulating more realistic, real-world scenarios?

4.4.2 Datasets

For a fair comparison with existing methodologies, we assembled an unpaired multi-modal dataset for abdominal multi-organ segmentation, consisting of 162 CT scans and 54 MR scans from AMOS [57] dataset. The focus is on segmenting 13 abdominal organs: spleen (SPL), right kidney (RKI), left kidney (LKI), gallbladder (GBL), esophagus (ESO), liver (LIV), stomach (STO), aorta (AOR), inferior vena cava (IVC), pancreas (PAN), right adrenal gland (RAD), left adrenal gland

(LAG), and duodenum (DUO). Following established protocols, distinct preprocessing methods were applied to CT and MR scans to address modality discrepancies. CT scans are clipped to the intensity window $[-275, 125]$, and then normalized to $[0, 1]$, while MR scans were resampled to $[1.5 \times 1.5 \times 2.0]$ mm^3, cropped to $96 \times 96 \times 96$ for training, and intensity histograms clipped by 0.5% before min-max normalization to [0, 1]. The Multi-Modality Whole Heart Segmentation Challenge [166] (MMWHS) dataset comprises 20 CT and 20 MR scans, collectively used for cardiac substructure segmentation, focusing on seven substructures: left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium of LV (MY), ascending aorta (AA), and pulmonary artery (PA). Preprocessing involves resampling scans to $[1.5\times1.5\times2.0]$ mm^3 resolution and normalizing intensities to $[0, 1]$ range, cropped to $96 \times 96 \times 96$ for training.

Balanced Data Splitting We utilized an equal ratio of CT to MR scans for the AMOS dataset, each with 54 scans. Of these, 35 scans are used for training and 19 for testing. Similarly, for the MMWHS dataset, each modality (CT or MR) consists of 20 scans, divided into 75% for training and 25% for testing.

Imbalanced Data Splitting In the AMOS dataset, we implemented ratios of 2:1 and 3:1 for CT to MR scans, thus increasing the number of CT scans to deviate from the balanced splitting. For the MMWHS dataset, building on previous cross-modality segmentation research on this dataset [73, 9], we employed MR as the auxiliary modality and CT as the target modality. This choice was driven by the superior soft tissue contrast provided by MR, which offers more detailed information for segmenting heart substructures. We divided the CT data randomly and evenly to conduct a two-fold cross-validation. Each training iteration involved 20 MRs and 10 CTs, simulating a scenario of data scarcity for the target modality.

4.4.3 Implementation Details

We utilized PyTorch 2.0 and MONAI 1.2 [16] to implement both our proposed method and the baselines for comparison. All models were trained from scratch on a server with NVIDIA A100 GPUs. Specifically, we assessed the performance of 3D UNet and SwinUNETR as backbones for our MulModSeg strategy. We employed the AdamW optimizer with a warm-up cosine scheduler with an initial learning rate of 10^{-3} and a weight decay of 10^{-4} for 1000 epochs training including the first 10 epochs for warmup. The last epoch's model is used for evaluation based on empirical experience. All hyperparameters are obtained through two-fold cross-validation over the training set unless otherwise specified. To avoid overfitting, on-the-fly data augmentation is applied, including random foreground and background patch sampling with a $1:1$ ratio and intensity shifting/scaling. A sum of Dice loss and Cross-entropy loss is used for training. For inference, an overlapping area ratio of 0.5 is applied via a sliding window strategy, and the Dice score is used for performance evaluation. We will release the source codes upon acceptance.

4.5 Results and Discussion

Method	Text Description	Avg. Dice↑ (CT)	Avg. Dice↑ (MR)
Vision-Only	-	82.50	81.91
One Hot [146]	-	86.64	84.50
BioBERT [66]	A {CT/MR} imaging of a [CLS].	86.62	84.59
MedCLIP [134]	A {CT/MR} imaging of a [CLS].	86.30	84.10
V1-CLIP	A photo of [CLS].	85.60	84.30
V2-CLIP	There is [CLS] in this {CT/MR}.	86.20	84.70
V3-CLIP (**Ours**)	A {CT/MR} imaging of a [CLS].	**87.14**	**85.33**

Table 8: Performance comparison of different text embeddings of MulModSeg on AMOS dataset with UNet backbone. V3-CLIP achieves the highest mean Dice scores for both CT and MR modalities. The **bolded** text represents the best performance. CT: computerized tomography, MR: magnetic resonance. The default Vison-Only model does not use text information and is trained with ALT.

4.5.1 Effectiveness of Different Text Embeddings (*Q1*)

To investigate the impact of text embeddings in addition to vision-only and their variants on segmentation performance, we conducted ablation studies using different embeddings, including One Hot, BioBERT, MedCLIP, and our V1, V2, and V3 -CLIP. Table 8 presents the results for AMOS dataset with a balanced split and UNet backbone. The findings demonstrate a clear advantage of using modality-conditioned text embeddings over conventional methods. Specifically, our proposed V3-CLIP embedding achieved the highest mean Dice scores of 87.14 for CT and 85.33 for MR, significantly outperforming other embeddings. *This improvement underscores the efficacy of leveraging modality-specific textual information to enhance feature representation and segmentation accuracy of the computer vision models*.

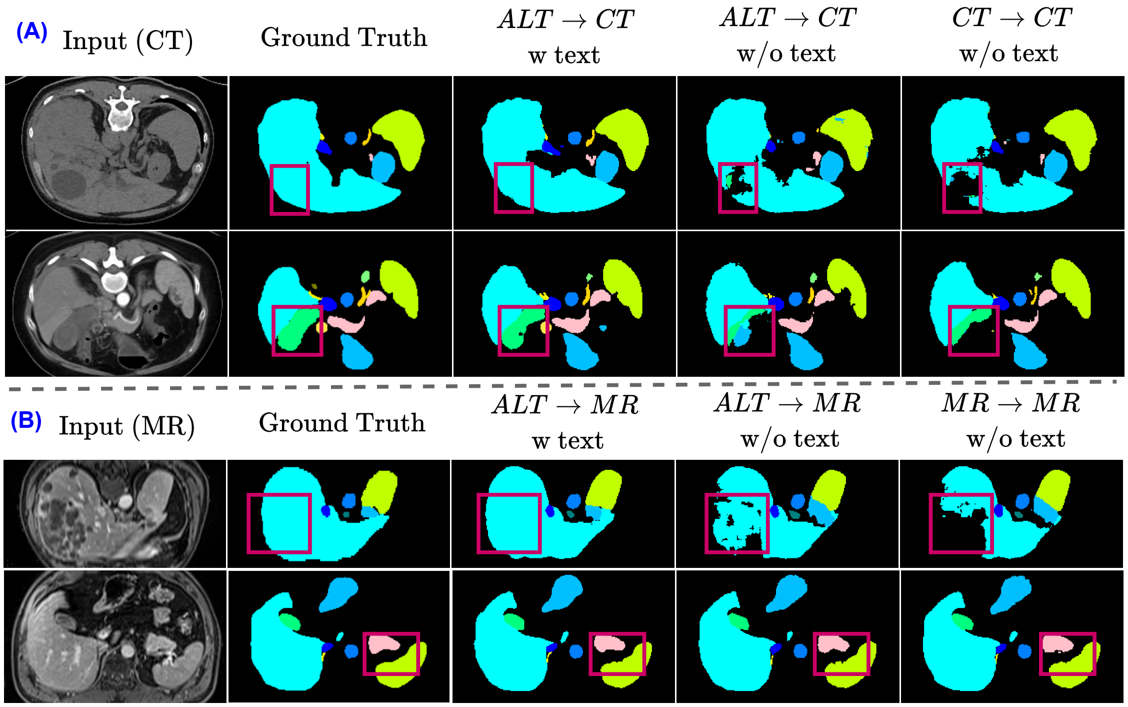
(A)
Input (CT)
Ground Truth
$ALT \rightarrow CT$ w text
$ALT \rightarrow CT$ w/o text
$CT \rightarrow CT$ w/o text
(B)
Input (MR)
Ground Truth
$ALT \rightarrow MR$ w text
$ALT \rightarrow MR$ w/o text
$MR \rightarrow MR$ w/o text

Figure 13: Visual comparison of segmentation results on AMOS dataset using the UNet backbone. Red boxes highlight areas where MulModSeg demonstrates improved predicted details compared to baselines.

4.5.2 Positive Impacts of Modality-Conditioned Text Embedding and ALT (*Q2*)

Table 9 presents Dice scores for unpaired multi-modal abdominal multi-organ segmentation using UNet and SwinUNETR backbones on AMOS dataset. It compares segmentation accuracy across various settings: without text embedding (w/o text) and with text embedding (w text), and for different scenarios including CT (for training) → CT (for testing), ALT → CT, MR → MR, and ALT → MR. For both UNet and SwinUNETR backbones, configurations with text embeddings show substantial improvements over those without. For instance, with the UNet backbone, the ALT + Text setup yields an average Dice score of 87.14 on the CT test set, compared to 82.50 without text embeddings, and an increase from 81.91 to 85.33 on the MR test set. The SwinUNETR backbone shows similar trends, further validating our approach. Additionally, the ALT procedure alone outperforms setups without ALT, highlighting its effectiveness. Similar enhancements of the details in predicted masks can be observed in visual comparison results (Figure 13).

Setting		(**UNet**) Cat. Dice (%) ↑													Avg.↑
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAD	LAG	DUO	
w/o text	CT→CT	92.01	93.62	93.56	76.81	69.53	95.51	88.02	92.26	86.31	78.72	61.87	62.61	65.22	81.23
	ALT→CT	92.90	94.19	94.31	77.02	71.26	95.62	88.87	93.38	86.86	80.91	63.82	64.95	68.42	82.50
	MR→MR	94.57	93.91	93.12	67.38	70.12	95.87	86.13	91.02	86.44	79.42	55.73	51.20	60.13	78.85
	ALT→MR	95.20	95.58	95.32	70.71	74.42	96.37	88.18	91.26	87.82	81.01	60.83	63.82	64.28	81.91
w text	**ALT→CT**	94.18	95.62	95.53	84.57	78.56	96.30	91.97	94.82	89.07	86.65	70.02	76.60	78.97	87.14
	ALT→MR	96.06	95.99	95.85	81.00	78.55	97.40	90.92	92.28	90.20	87.03	67.38	66.32	70.30	85.33

Setting		(**SwinUNETR**) Cat. Dice (%) ↑													Avg.↑
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAD	LAG	DUO	
w/o text	CT→CT	94.02	95.39	95.22	82.41	77.57	95.89	88.76	94.50	88.25	84.61	64.33	71.17	73.67	85.06
	ALT→CT	94.49	95.03	95.11	79.39	78.35	95.98	88.00	94.05	88.17	88.17	62.95	72.44	73.18	85.02
	MR→MR	95.73	95.06	94.75	69.94	75.30	96.24	84.58	90.35	87.34	81.71	56.97	57.47	63.67	80.70
	ALT→MR	95.79	95.81	95.42	67.03	76.24	96.28	86.19	91.83	87.58	83.02	62.17	65.22	64.51	82.08
w text	**ALT→CT**	95.32	95.75	95.43	84.24	77.56	96.49	90.04	94.05	88.42	84.91	69.03	72.92	75.81	86.15
	ALT→MR	96.00	95.84	95.70	68.80	76.30	96.57	88.31	91.32	89.38	84.41	65.52	67.23	68.43	83.37

Table 9: Dice scores on AMOS dataset with balanced data splitting of MulModSeg using UNet and SwinUNETR backbones. Results are shown for settings without text embedding (w/o text) and with text embedding (w text) across various training and testing scenarios. Green and blue color represent the best performance for the CT and MR testing set, respectively.

Table 10 illustrates the significant performance gains achieved by the MulModSeg framework using modality-conditioned text embeddings and the ALT procedure for unpaired multi-modal cardiac substructure segmentation within MMWHS dataset. For the UNet backbone, the ALT + Text configuration achieves an average Dice score of 91.67 on the CT test set, compared to 90.41 without text embeddings, and an increase from 82.82 to 85.15 on the MR test set. Similar trends are observed with the SwinUNETR backbone. Additionally, the ALT procedure alone shows improved performance over non-ALT configurations. *These results confirm that combining text embeddings with ALT significantly enhances the model's generalization and segmentation accuracy across modalities. This demonstrates the robustness and effectiveness of the MulModSeg strategy in diverse cardiac and abdominal multi-organ segmentation tasks.*

Setting		Avg.↑	Dice of Substructures of Heart (UNet) ↑						
			MY	LA	LV	RA	RV	AA	PA
w/o text	CT→CT	90.55	90.54	94.25	88.75	87.12	91.53	95.87	85.81
	ALT→CT	90.41	90.80	94.56	88.82	85.37	92.00	95.10	86.24
	MR→MR	81.04	81.11	85.92	74.72	83.61	87.32	74.63	79.99
	ALT→MR	82.82	81.00	87.03	79.03	85.62	88.06	76.90	82.22
w text	**ALT→CT**	91.67	91.40	95.28	90.82	88.57	91.87	96.50	87.23
	ALT→MR	85.15	83.55	89.29	83.27	87.45	88.30	80.52	83.67

Setting		Avg.↑	Dice of Substructures of Heart (SwinUNETR) ↑						
			MY	LA	LV	RA	RV	AA	PA
w/o text	CT→CT	90.87	90.53	95.01	90.49	86.42	90.89	95.95	86.79
	ALT→CT	91.11	90.98	94.84	90.26	87.35	92.22	96.07	86.27
	MR→MR	83.07	83.17	90.39	81.05	84.44	87.67	76.41	78.36
	ALT→MR	83.33	83.07	88.43	81.52	85.88	86.79	76.67	80.94
w text	**ALT→CT**	91.44	91.10	95.13	90.25	88.00	91.89	96.33	87.36
	ALT→MR	83.85	82.48	90.22	82.49	85.32	88.39	77.24	80.79

Table 10: Dice scores with UNet and SwinUNETR backbone for MMWHS dataset with balanced data splitting. Green and blue color represent the best performance for the CT and MR testing set, respectively.

4.5.3 Comparison with State-of-the-Art Methods (*Q3*)

To benchmark the performance of our MulModSeg framework, we compared it against several state-of-the-art methods for cross-modality medical image segmentation on MMWHS dataset in an unbalanced data splitting setting, focusing on the target modality (CT) using both UNet and SwinUNETR backbones. Table 11 shows that MulModSeg consistently outperforms existing methods, achieving the highest average Dice scores of 92.72 for the UNet backbone and 93.31 for the SwinUNETR backbone. Specifically, MulModSeg demonstrated substantial improvements in segmenting cardiac substructures such as the left ventricle (LV), left atrium (LA), and ascending aorta (AA), with Dice scores of 92.73, 93.42, and 94.80 for UNet, and 93.50 and 91.59 for LV and RA with SwinUNETR, respectively. These results highlight the robustness and precision of our method, attributed to the innovative use of modality-conditioned text embeddings and the ALT procedure, which enhance the model's ability to leverage information from unpaired multi-modal datasets. *This allows MulModSeg to achieve high segmentation accuracy without requiring paired images or extensive architectural modifications, establishing it as a state-of-the-art solution for multi-modal medical image segmentation.*

4.5.4 Impact of Varying CT to MR Scan Ratios (*Q4*)

We also explored the impact of varying the ratio of CT to MR scans on the segmentation performance to simulate more realistic clinical scenarios. Table 12 presents the Dice scores for different CT:MR ratios (1:1, 2:1, and 3:1) on AMOS dataset using the UNet backbone. The results indicate that increasing the number of CT scans relative to MR scans generally enhances segmentation performance. For instance, the average Dice score for the 3:1 CT:MR ratio was 88.92, compared to 87.14 for the 1:1 ratio. *This trend is consistent across both CT and MR testing sets, suggesting that*

Model	Avg. ↑	Dice of Substructures of Heart (UNet) ↑						
		MY	LA	LV	RA	RV	AA	PA
Baseline [73]	87.06	87.02	89.22	90.86	83.86	84.60	92.52	81.34
Fine-tune [73]	87.69	87.16	90.40	90.79	84.43	85.26	92.74	83.05
Joint-training[73]	87.43	86.65	90.76	91.23	82.78	84.92	93.02	82.66
X-shape [124]	87.67	87.19	89.79	90.94	85.51	84.44	93.43	82.40
Zhang *et al.*[151]	88.50	87.81	91.12	91.34	85.14	86.31	94.30	83.42
Li *et al.*[73]	90.12	89.34	91.90	92.67	87.47	88.14	**95.95**	85.38
Bastico *et al.*[9]	90.77	90.06	92.68	**93.77**	88.22	90.85	94.70	84.52
Ours	**92.72**	**92.32**	**92.73**	89.85	**93.42**	**94.80**	95.48	**90.46**

Model	Avg. ↑	Dice of Substructures of Heart (SwinUNETR) ↑						
		MY	LA	LV	RA	RV	AA	PA
Baseline[151]	85.32	84.79	88.12	89.80	81.04	84.08	77.58	76.36
Fine-tune[151]	86.09	82.57	90.03	87.92	83.29	85.51	90.28	83.06
Joint-training[151]	87.99	86.25	92.07	91.97	85.35	87.98	89.55	82.76
Bastico *et al.*[9]	89.33	88.13	91.64	**92.39**	86.36	89.33	93.39	84.05
Ours	**93.31**	**93.15**	**93.50**	91.59	**92.20**	**95.63**	**95.83**	**91.27**

Table 11: Quantitative comparison with other methods for cross-modality medical image segmentation on the target modality (CT). All the techniques have the same UNet [124] and SwinUNETR [47] baseline, are trained using 20 MRs and 10 CTs and are evaluated on the test set of MMWHS dataset. The mean Dice score is reported, as well as the ones for all the heart substructures.

our approach can effectively leverage the availability of more CT data to improve segmentation accuracy while maintaining robust performance across modalities. Despite only increasing the number of CTs, we believe that utilizing more MRs in our MulModSeg architecture can achieve a similar enhancement in performance for both modalities.

4.5.5 Ablation Study

The ablation study (Table 13) confirms the contribution of each component of our MulModSeg strategy. The combination of text embedding and ALT significantly boosts segmentation performance, with the highest mean Dice scores observed for both CT (87.14) and MR (85.33) testing sets. Furthermore, our model maintains a balance between performance and complexity, as indi-

Ratio of CT:MR		AMOS-CT Testing Cat. Dice ↑													Avg.↑
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAD	LAG	DUO	
MulModSeg	1:1	94.18	95.62	95.53	84.57	78.56	96.30	91.97	94.82	89.07	86.65	70.02	76.60	78.97	87.14
	2:1	95.23	95.72	96.01	85.02	80.07	96.40	92.04	95.07	89.91	86.65	71.93	76.75	80.80	87.82
	3:1	95.72	96.14	96.22	86.69	83.85	97.22	93.77	95.39	90.29	87.99	72.48	79.30	80.89	88.92

Ratio of CT:MR		AMOS-MR Testing Cat. Dice ↑													Avg.↑
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAD	LAG	DUO	
MulModSeg	1:1	96.06	95.99	95.85	81.00	78.55	97.40	90.92	92.28	90.20	87.03	67.38	66.32	70.30	85.33
	2:1	96.42	95.94	95.81	76.11	79.30	97.50	90.93	92.54	90.83	85.84	64.24	72.01	72.53	85.38
	3:1	96.40	96.06	96.18	77.14	79.56	97.58	91.17	92.15	90.59	86.62	64.28	71.42	72.28	85.49

Table 12: Dice scores for abdominal multi-organ segmentation on AMOS dataset using the UNet backbone, with varying CT to MR scan ratios (1:1, 2:1, and 3:1). Results show the impact of different data ratios on segmentation performance.

cated by the model parameters and inference time (Table 14).

Text Emb.	Training	Avg. Dice↑ (CT)	Avg. Dice↑ (MR)
●	*ALT*	**87.14**	**85.33**
○	*ALT*	82.50	81.91
●	*CT*	86.40	69.72
●	*MR*	60.53	82.70
○	*CT*	81.23	55.45
○	*MR*	46.64	78.85

Table 13: Ablation study for the MulModSeg strategy. ●: with, ○: without. The **bolded** text represents the best performance in AMOS testing sets with UNet backbone.

4.6 Conclusion

In this chapter, we introduced MulModSeg, a multi-modal segmentation strategy enhancing CT and MR medical image segmentation. MulModSeg leverages modality-conditioned text embeddings and an alternating training (ALT) procedure, integrating modality-specific information into existing encoder-decoder frameworks without significant architectural changes. Extensive experiments showed that MulModSeg significantly improves segmentation accuracy and robustness over state-of-the-art methods, achieving higher Dice scores for abdominal multi-organ and cardiac substructure segmentation tasks. The method's adaptability to different imaging modalities and

Text Emb.	Backbone	Prams. (M)	Time (s)
●	UNet	19.4	3.272
○	UNet	19.1	2.402
●	SwinUNETR	62.6	4.197
○	SwinUNETR	62.2	3.387

Table 14: Model parameters (in millions) and inference time per case (in seconds) for AMOS CT testing set. Comparison of UNet and SwinUNETR backbones with and without text embedding. ●: with, ○: without.

balanced/imbalanced training scenarios across unpaired datasets ensure practical clinical application. This flexible strategy improves diagnostic accuracy and medical image analysis, with future work aiming to extend its application to other imaging modalities and clinical scenarios.

CHAPTER 5 SUMMARY AND FUTURE OUTLOOK

5.1 Summary

This dissertation has embarked on a transformative journey in the realm of medical image segmentation, aiming to overcome the inherent limitations of CNNs through the adoption of transformer-based models. The three significant contributions presented in this dissertation, namely FocalUNETR, SwinAttUNet, and MulModSeg.

FocalUNETR, introduced in Chapter 2, emerged as a pioneering 2D transformer-based model meticulously designed to address the challenges inherent in medical image segmentation, particularly in the domain of CT scans. By incorporating focal self-attention mechanisms, FocalUNETR not only improved segmentation accuracy but also redefined the standards of precision in medical image analysis. This innovation marked a crucial departure from conventional CNN approaches, ushering in a new era of transformative possibilities. Despite FocalUNETR's success, its application is currently limited to 2D-based single-organ segmentation. Owing to the challenges in designing an efficient 3D version of the focal SA, a viable 3D-based multi-organ segmentation approach remains unachievable.

In Chapter 3, SwinAttUNet, further extended the boundaries of this research into the 3D realm of multi-organ segmentation. This model, based on transformer principles, efficiently processed intricate 3D medical images while preserving essential spatial relationships. Its remarkable performance surpassed that of conventional methods, underscoring its potential to revolutionize 3D medical imaging and analysis. However, it's important to acknowledge that SwinAttUNet's training involves a relatively smaller number of medical images. Given the abundance of labeled natural images, there's a pressing need to explore more effective ways to leverage these resources to further advance the field of medical image segmentation.

Chapter 4 presented MulModSeg to address the challenge of the emerging need for robust segmentation across diverse medical imaging modalities. It incorporates two key innovations: a modality-specific text embedding via CLIP model that adds modality awareness to existing segmentation frameworks without significant computational overhead or structural modifications, and an alternating training method that facilitates the integration of essential features from unpaired images. Applied to both FCN and Transformer-based models with extensive experiments, MulModSeg demonstrates superior performance in segmenting abdominal multi-organ and cardiac substructures compared to the existing strategies.

In summary, these contributions signify a promising future in medical image segmentation. The integration of advanced deep learning methods i.e., transformer-based models promises to unlock new frontiers of accuracy and efficiency. As we conclude this dissertation, we recognize the profound impact it is poised to have on the future of medical image analysis, ultimately contributing to enhanced patient care and well-being on a global scale.

5.2 Future Outlook

The future outlook for medical image segmentation, building upon the foundations laid by the three papers presented in this dissertation, holds tremendous promise and aligns with the imperative to address three key areas in the field.

Firstly, the development of Universal Medical Image Segmentation Models is pivotal. Existing models often struggle when applied to new datasets containing previously unseen organs, underscoring the need for models that generalize across diverse medical imaging scenarios. The groundwork laid by FocalUNETR and SwinAttUNet showcases the potential of transformer-based models in achieving this universality. Future research should focus on creating versatile models that adapt seamlessly to various medical imaging contexts, reducing the need for dataset-specific architectures and expediting the deployment of accurate segmentation solutions.

Secondly, the integration of Multimodal Learning for Medical Image Segmentation offers a compelling solution to the challenges posed by the diverse range of medical image modalities. Incorporating robust language features alongside visual data, as exemplified in the MulModSeg architecture, can enhance the reliability of segmentation results. Future research should explore innovative strategies for fusing both visual and language modalities, leveraging the strengths of each to improve the overall robustness of segmentation models. This multimodal approach holds the potential to revolutionize medical image segmentation, making it more adaptable, interpretable, and capable of handling the intricacies of modern healthcare imaging.

Lastly, the concept of Foundation Models for Medical Image Segmentation, inspired by the success in natural language processing and natural image domains, presents an exciting avenue for exploration. These models, trained on vast and diverse datasets with a consistent learning objective, have the potential to learn shared concepts that prove robust during inference across different datasets and modalities. Future work should delve deeper into the creation of specialized foundation models tailored to the medical domain. These models can serve as the cornerstone for more resilient and accurate medical image segmentation, setting new standards for precision and reliability.

In summary, the future of medical image segmentation is poised to advance on multiple fronts, guided by the principles of universality, multimodal learning, and foundation models. Researchers and practitioners have the opportunity to shape the future of healthcare diagnostics and treatment planning, ultimately benefiting patients worldwide.

APPENDIX A: DATA PROCESSING AND VISUALIZATION

The combination of these data processing and visualization techniques facilitated the effective analysis and segmentation of medical images, contributing to the advancements in automated medical image analysis presented in this dissertation.

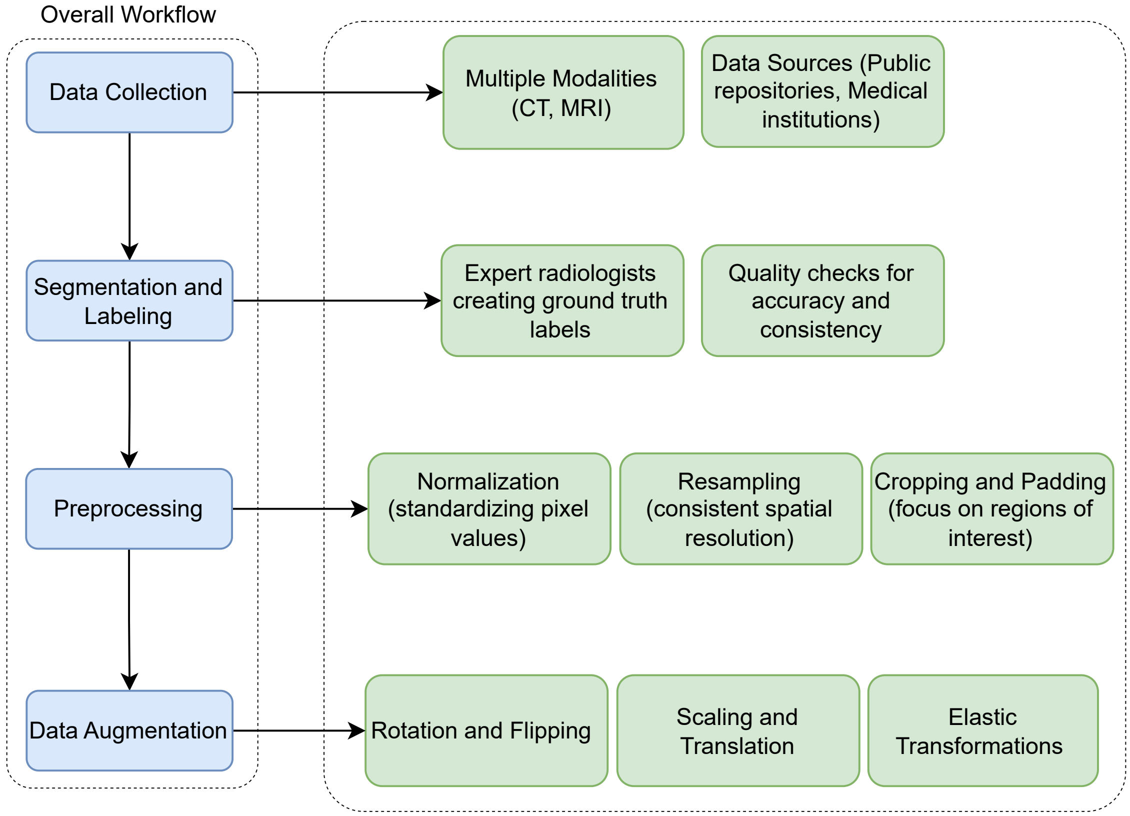
Overall Workflow
Data Collection
Segmentation and Labeling
Preprocessing
Data Augmentation
Multiple Modalities (CT, MRI)
Data Sources (Public repositories, Medical institutions)
Expert radiologists creating ground truth labels
Quality checks for accuracy and consistency
Normalization (standardizing pixel values)
Resampling (consistent spatial resolution)
Cropping and Padding (focus on regions of interest)
Rotation and Flipping
Scaling and Translation
Elastic Transformations

Figure 14: A workflow for medical image data processing for developing deep learning models.

Data Processing

The data processing workflow applied in this dissertation focused primarily on the preparation and transformation of medical imaging data for effective segmentation and analysis and is summarized in Figure 14. The steps involved in data processing are as follows:

Data Collection

Medical imaging data were collected from multiple modalities, including Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). These datasets were sourced from publicly available medical image repositories and through collaborations with medical institutions.

Segmentation and Labeling

Ground truth labels for segmentation tasks were created by expert radiologists. These labels served as the benchmark for training and evaluating the segmentation models.

Preprocessing

To train deep learning models we have to perform some preprocessing for the collected medical images. It involved several steps to ensure the data was in an optimal format for analysis:

- **Normalization**: Image intensities were normalized to standardize the pixel values across different scans, which helps in reducing the variability due to different imaging conditions.
- **Resampling**: Images were resampled to a consistent spatial resolution to ensure uniformity in the analysis and to facilitate the integration of data from different sources.
- **Cropping and Padding**: Regions of interest were cropped to focus on specific anatomical areas, and padding was applied to maintain a consistent input size for the neural networks.

Data Augmentation

To enhance the robustness of the model and to prevent overfitting, data augmentation techniques were employed:

- **Rotation and Flipping**: Random rotations and flips were applied to the images to simulate different orientations and improve the model's generalization.

- **Scaling and Translation**: Variations in scale and translation were introduced to make the model invariant to size and positional changes.

- **Elastic Transformations**: Elastic deformations were used to mimic realistic variations in anatomical structures.

Data Visualization

Effective visualization techniques were critical for the interpretation and validation of the segmentation results. The following visualization methods were utilized:

Slice-wise Visualization

For volumetric data, slice-wise visualization allowed the inspection of individual 2D slices from the 3D volume. This method was particularly useful for verifying the accuracy of segmentation boundaries on a per-slice basis.

3D Volume Rendering

3D volume rendering provided a comprehensive view of the segmented regions within the entire volumetric scan. This technique enabled the visualization of complex anatomical structures and the spatial relationships between different organs and tissues.

Overlay Visualization

Segmented regions were overlaid on the original images to visually assess the accuracy of the segmentation. Different colors were used to distinguish between various anatomical structures, making it easier to identify segmentation errors.

APPENDIX B: PUBLICATIONS

Author Publications

[1] **Chengyin Li**, Yao Qiang, Rafi Ibn Sultan, Hassan Bagher-Ebadian, Prashant Khanduri, Indrin J. Chetty, and Dongxiao Zhu. "FocalUNETR: A Focal Transformer for Boundary-Aware Prostate Segmentation Using CT Images." *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), pp. 592-602. Cham: Springer Nature Switzerland, 2023.

[2] **Chengyin Li**, Hassan Bagher-Ebadian, Rafi Ibn Sultan, Mohamed Elshaikh, Benjamin Movsas, Dongxiao Zhu, and Indrin J. Chetty. "A New Architecture Combining Convolutional and Transformer-based Networks for Automatic 3D Multi-organ Segmentation on CT Images." *Medical Physics* 50, no. 11, 2023: 6990-7002.

[3] **Chengyin Li**, Zheng Dong, Nathan Fisher, and Dongxiao Zhu. "Coupling User Preference with External Rewards to Enable Driver-centered and Resource-aware EV Charging Recommendation." *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (*ECML-PKDD, Oral*), pp. 3-19. Cham: Springer Nature Switzerland, 2022.

[4] **Chengyin Li**, Rhea E. Sullivan, Dongxiao Zhu, and Steven D. Hicks. "Putting The ’mi’ in Omics: Discovering miRNA Biomarkers for Pediatric Precision Care." *Pediatric research*, 2023.

[5] **Chengyin Li**, Rafi Ibn Sultan, Hassan Bagher-Ebadian, Yao Qiang, Kundan S. Thind, Dongxiao Zhu, and Indrin J. Chetty. "On the Implementation and Evaluation of Loss Functions for Robust Multiple Anatomy Segmentation on CT Images." *International Conference on the use of Computers in Radiation Therapy (ICCR)*, 2024.

[6] Xin Li, Deng Pan, **Chengyin Li**, Yao Qiang, and Dongxiao Zhu. "Negative Flux Aggregation to Estimate Feature Attributions." *In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 446-454, 2023.

[7] Prashant Khanduri, **Chengyin Li**, Rafi Ibn Sultan, Yao Qiang, Joerg Kliewer, and Dongxiao Zhu. "Proximal Compositional Optimization for Distributionally Robust Learning." *In The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[8] Yao Qiang, **Chengyin Li**, Marco Brocanelli, and Dongxiao Zhu. "Counterfactual Interpolation Augmentation (CIA): A Unified Approach to Enhance Fairness and Explainability of DNN." In *IJCAI*, pp. 732-739. 2022.

[9] Yao Qiang, Deng Pan, **Chengyin Li**, Xin Li, Rhongho Jang, and Dongxiao Zhu. "AttCAT: Explaining transformers via attentive class activation tokens." *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[10] Xin Li, Yao Qiang, **Chengyin Li**, Sijia Liu, and Dongxiao Zhu. "Saliency-guided Adversarial Training for Learning Generalizable Features with Applications to Medical Imaging Classification System". *ICML New Frontiers in Adversarial Machine Learning Workshop*, 2022.

[11] Xin Li, **Chengyin Li**, and Dongxiao Zhu. "COVID-MobileXpert: On-device COVID-19 Patient Triage and Follow-up using Chest X-rays." *In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1063-1067. IEEE, 2020.

Under Review

[1] **Chengyin Li**, Hui Zhu, Rafi Ibn Sultan, Indrin J. Chetty, Kundan Thind, and Dongxiao Zhu. "MulModSeg: Enhancing Unpaired Multi-Modal Segmentation for Medical Images."

[2] **Chengyin Li**, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin J. Chetty, and Dongxiao Zhu. "Auto-prompting sam for Mobile Friendly 3D Medical Image Segmentation." arXiv preprint arXiv:2308.14936, 2023.

[3] Yao Qiang, **Chengyin Li**, Prashant Khanduri, and Dongxiao Zhu. "Fairness-aware Vision Transformer via Debiased Self-attention." arXiv preprint arXiv:2301.13803, 2023.

[4] Yao Qiang, **Chengyin Li**, Prashant Khanduri, and Dongxiao Zhu. "Interpretability-aware Vision Transformer." arXiv preprint arXiv:2309.08035, 2023.

APPENDIX C: DECLARATION OF USING AIGC

For the role of AIGC in this dissertation, I used ChatGPT solely for grammar checking, proofreading, and revising content I have already written. I did not use any AIGC tools to generate creative content.

REFERENCES

[1] M. A. Abdou. Literature review: Efficient deep neural networks techniques for medical image analysis. *Neural Computing and Applications*, 34(8):5791–5812, 2022.

[2] F. Altaf, S. M. Islam, N. Akhtar, and N. K. Janjua. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019.

[3] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[4] S. G. Armato III, H. Huisman, K. Drukker, L. Hadjiiski, J. S. Kirby, N. Petrick, G. Redmond, M. L. Giger, K. Cha, A. Mamonov, et al. Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501–044501, 2018.

[5] A. Balagopal, S. Kazemifar, D. Nguyen, M.-H. Lin, R. Hannan, A. Owrangi, and S. Jiang. Fully automated organ segmentation in male pelvic ct images. *Physics in Medicine & Biology*, 63(24):245015, 2018.

[6] A. Balagopal, H. Morgan, M. Dohopolski, R. Timmerman, J. Shan, D. F. Heitjan, W. Liu, D. Nguyen, R. Hannan, A. Garant, et al. Psa-net: Deep learning–based physician style–aware segmentation network for postoperative prostate cancer clinical target volumes. *Artificial Intelligence in Medicine*, 121:102195, 2021.

[7] A. Balagopal, D. Nguyen, H. Morgan, Y. Weng, M. Dohopolski, M.-H. Lin, A. S. Barkousaraie, Y. Gonzalez, A. Garant, N. Desai, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative

prostate cancer radiotherapy. *Medical image analysis*, 72:102101, 2021.

[8] A. M. Barhoom, M. R. J. Al-Hiealy, and S. S. Abu-Naser. Deep learning-xception algorithm for upper bone abnormalities classification. *Journal of Theoretical and Applied Information Technology*, 100(23):6986–6997, 2022.

[9] M. Bastico, D. Ryckelynck, L. Corté, Y. Tillier, and E. Decencière. A simple and robust framework for cross-modality medical image segmentation applied to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4128–4138, 2023.

[10] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

[11] A. L. Breto, B. Spieler, O. Zavala-Romero, M. Alhusseini, N. V. Patel, D. A. Asher, I. R. Xu, J. B. Baikovitz, E. A. Mellon, J. C. Ford, et al. Deep learning for per-fraction automatic segmentation of gross tumor volume (gtv) and organs at risk (oars) in adaptive radiotherapy of cervical cancer. *Frontiers in oncology*, 12:854349, 2022.

[12] T. M. Buzug. Computed tomography. In *Springer handbook of medical technology*, pages 311–342. Springer, 2011.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[15] C. Cardenas, J. Yang, and B. Anderson. Court le, brock kb. advances in auto-segmentation. *Semin Radiat Oncol*, 29(3):185–197, 2019.

[16] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[17] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2*, pages 3–13. Springer, 2017.

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

[19] C. Chen and G. Zheng. Fully automatic segmentation of ap pelvis x-rays via random forest regression with efficient feature selection and hierarchical sparse shape composition. *Computer vision and image understanding*, 126:1–10, 2014.

[20] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[21] Y. Chen, C. Liu, W. Huang, S. Cheng, R. Arcucci, and Z. Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023.

[22] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang. CrDoCo: Pixel-level Domain Transfer with Cross-Domain Consistency, Jan. 2020. arXiv:2001.03182 [cs].

[23] 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*, pages 424–432. Springer, 2016.

[24] A. Conneau and G. Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[25] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[26] S. Cui, H.-H. Tseng, J. Pakela, R. K. Ten Haken, and I. El Naqa. Introduction to machine and deep learning for medical physicists. *Medical physics*, 47(5):e127–e147, 2020.

[27] O. Dalmaz, M. Yurt, and T. Çukur. ResViT: Residual vision transformers for multi-modal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, Oct. 2022. arXiv:2106.16031 [cs, eess].

[28] A. V. D'Amico, R. Whittington, S. B. Malkowicz, D. Schultz, K. Blank, G. A. Broderick, J. E. Tomaszewski, A. A. Renshaw, I. Kaplan, C. J. Beard, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Jama*, 280(11):969–974, 1998.

[29] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on medical imaging*, 38(5):1116–1126, 2018.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth

16x16 words: Transformers for image recognition at scale. *ICLR*, abs/2010.11929, 2021.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[32] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020.

[33] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019.

[34] J. S. Dramsch, A. N. Christensen, C. MacBeth, and M. Lüthje. Deep unsupervised 4-d seismic 3-d time-shift estimation with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[35] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 179–187. Springer, 2016.

[36] I. El Naqa and S. Das. The role of machine and deep learning in modern medical physics. 2020.

[37] H. L. Elghazy and M. W. Fakhr. Multi-modal multi-stream unet model for liver segmentation. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0028–0033. IEEE, 2021.

[38] 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*,

4(1):5, 2021.

[39] M. Feng, G. Valdes, N. Dixit, and T. D. Solberg. Machine learning in radiation oncology: opportunities, requirements, and needs. *Frontiers in oncology*, 8:110, 2018.

[40] Y. Gao, Y. Shao, J. Lian, A. Z. Wang, R. C. Chen, and D. Shen. Accurate segmentation of ct male pelvic organs via regression-based deformable models and multi-task random forests. *IEEE transactions on medical imaging*, 35(6):1532–1543, 2016.

[41] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022.

[42] Y. Gao, M. Zhou, and D. N. Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer, 2021.

[43] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi. Joint classification-regression forests for spatially structured multi-object segmentation. In *European conference on computer vision*, pages 870–881. Springer, 2012.

[44] E. Gudmundsson, C. M. Straus, F. Li, and S. G. Armato III. Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion. *Journal of Medical Imaging*, 7(1):012705–012705, 2020.

[45] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li. Medical Image Segmentation Based on Multi-Modal Convolutional Neural Network: Study on Image Fusion Schemes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 903–907, Apr. 2018. arXiv:1711.00049 [cs].

[46] L. Hadjiiski, K. Cha, H.-P. Chan, K. Drukker, L. Morra, J. J. Näppi, B. Sahiner, H. Yoshida, Q. Chen, T. M. Deserno, et al. Aapm task group report 273: Recommendations on best practices for ai and machine learning for computer-aided diagnosis in medical imaging. *Medical Physics*, 50(2):e1–e24, 2023.

[47] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[48] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[49] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen. Transformers in medical image analysis: A review. *arXiv preprint arXiv:2202.12165*, 2022.

[50] K. He, C. Lian, B. Zhang, X. Zhang, X. Cao, D. Nie, Y. Gao, J. Zhang, and D. Shen. Hf-unet: learning hierarchically inter-task relevance in multi-task u-net for accurate prostate segmentation in ct images. *IEEE Transactions on Medical Imaging*, 40(8):2118–2128, 2021.

[51] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[52] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

[53] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network, Mar. 2015. arXiv:1503.02531 [cs, stat].

[54] Y. Interian, V. Rideout, V. P. Kearney, E. Gennatas, O. Morin, J. Cheung, T. Solberg, and G. Valdes. Deep nets vs expert designed features in medical physics: an imrt qa case study. *Medical physics*, 45(6):2672–2680, 2018.

[55] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, pages 118–132. Springer, 2021.

[56] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[57] Y. Ji, H. Bai, J. Yang, C. Ge, Y. Zhu, R. Zhang, Z. Li, L. Zhang, W. Ma, X. Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

[58] J. Jiang, S. Elguindi, S. L. Berry, I. Onochie, L. Cervino, J. O. Deasy, and H. Veeraraghavan. Nested-block self-attention multiple resolution residual network for multi-organ segmentation from ct. *Medical Physics*, 2022.

[59] J. Jiang, A. Rimner, J. O. Deasy, and H. Veeraraghavan. Unpaired cross-modality educed distillation (cmedl) for medical image segmentation. *IEEE Transactions on medical imaging*, 41(5):1057–1068, 2021.

[60] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets. Medical image segmentation using deep neural networks with pre-trained encoders. *Deep learning applications*, pages 39–52, 2020.

[61] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

[62] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owrangi. Segmentation of the prostate and organs at risk in male pelvic ct images using deep learning. *Biomedical Physics & Engineering Express*, 4(5):055003, 2018.

[63] V. Kearney, J. W. Chan, T. Wang, A. Perry, S. S. Yom, and T. D. Solberg. Attention-enabled 3d boosted convolutional neural networks for semantic ct segmentation using deep supervision. *Physics in Medicine & Biology*, 64(13):135001, 2019.

[64] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pages 285–296. PMLR, 2019.

[65] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

[66] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[67] Y. Lei, X. Dong, Z. Tian, Y. Liu, S. Tian, T. Wang, X. Jiang, P. Patel, A. B. Jani, H. Mao, et al. Ct prostate segmentation based on synthetic mri-aided deep attention fully convolution network. *Medical physics*, 47(2):530–540, 2020.

[68] L. Lenchik, L. Heacock, A. A. Weaver, R. D. Boutin, T. S. Cook, J. Itri, C. G. Filippi, R. P. Gullapalli, J. Lee, M. Zagurovskaya, et al. Automated segmentation of tissues using ct and mri: a systematic review. *Academic radiology*, 26(12):1695–1706, 2019.

[69] C. Li, H. Bagher-Ebadian, V. Goddla, I. J. Chetty, and D. Zhu. Focalunetr: A focal transformer for boundary-aware segmentation of ct images. *ArXiv*, abs/2210.03189, 2022.

[70] C. Li, H. Bagher-Ebadian, R. I. Sultan, M. Elshaikh, B. Movsas, D. Zhu, and I. J. Chetty. A new architecture combining convolutional and transformer-based networks for automatic 3d multi-organ segmentation on ct images. *Medical physics*, 2023.

[71] C. Li, Y. Qiang, R. I. Sultan, H. Bagher-Ebadian, P. Khanduri, I. J. Chetty, and D. Zhu. Focalunetr: A focal transformer for boundary-aware prostate segmentation using ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer, 2023.

[72] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, page 102762, 2023.

[73] K. Li, L. Yu, S. Wang, and P.-A. Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 775–783, 2020.

[74] K. Li, L. Yu, S. Wang, and P.-A. Heng. Towards Cross-modality Medical Image Segmentation with Online Mutual Knowledge Distillation, Oct. 2020. arXiv:2010.01532 [cs, eess].

[75] X. Li, H. Bagher-Ebadian, S. Gardner, J. Kim, M. Elshaikh, B. Movsas, D. Zhu, and I. J. Chetty. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. *Medical Physics*, 2022.

[76] X. Li, H. Bagher-Ebadian, S. Gardner, J. Kim, M. Elshaikh, B. Movsas, D. Zhu, and I. J. Chetty. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. *Medical physics*, 50(1):311–322, 2023.

[77] X. Liang, J.-E. Bibault, T. Leroy, A. Escande, W. Zhao, Y. Chen, M. K. Buyyounouski, S. L. Hancock, H. Bagshaw, and L. Xing. Automated contour propagation of the prostate from pct to cbct images via deep unsupervised learning. *Medical physics*, 48(4):1764–1770, 2021.

[78] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 2022.

[79] L. Lin, Z. Wang, J. Wu, Y. Huang, J. Lyu, P. Cheng, J. Wu, and X. Tang. Bsda-net: A boundary shape and distance aware joint learning framework for segmenting and classifying octa images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 65–75. Springer, 2021.

[80] C. Liu, S. Cheng, C. Chen, M. Qiao, W. Zhang, A. Shah, W. Bai, and R. Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2023.

[81] C. Liu, S. J. Gardner, N. Wen, M. A. Elshaikh, F. Siddiqui, B. Movsas, and I. J. Chetty. Automatic segmentation of the prostate on ct images using deep neural networks (dnn). *International Journal of Radiation Oncology* Biology* Physics*, 104(4):924–932, 2019.

[82] H. Liu, Y. Zhuang, E. Song, X. Xu, G. Ma, C. Cetinkaya, and C.-C. Hung. A modality-

collaborative convolution and transformer hybrid network for unpaired multi-modal medical image segmentation with limited annotations. *Medical Physics*, 2023.

[83] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[84] Q. Liu, Q. Dou, L. Yu, and P. A. Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9):2713–2724, 2020.

[85] 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*, pages 10012–10022, 2021.

[86] 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, Aug. 2021. arXiv:2103.14030 [cs].

[87] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[88] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

[89] J. Ma, Z. Wei, Y. Zhang, Y. Wang, R. Lv, C. Zhu, C. Gaoxiang, J. Liu, C. Peng, L. Wang, et al. How distance transform maps boost segmentation cnns: an empirical study. In *Medical Imaging with Deep Learning*, pages 479–492. PMLR, 2020.

[90] L. Ma, R. Guo, G. Zhang, D. M. Schuster, and B. Fei. A combined learning algorithm for prostate segmentation on 3d ct images. *Medical physics*, 44(11):5768–5781, 2017.

[91] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, W. Enbeyle, et al. Deep neural networks for medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

[92] Z. Marinov, S. Reiß, D. Kersting, J. Kleesiek, and R. Stiefelhagen. Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023.

[93] P. E. McKnight and J. Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[94] R. Metzner, A. Eggert, D. van Dusschoten, D. Pflugfelder, S. Gerth, U. Schurr, N. Uhlmann, and S. Jahnke. Direct comparison of mri and x-ray ct technologies for 3d imaging of root systems in soil: potential and challenges for root trait quantification. *Plant methods*, 11:1–11, 2015.

[95] 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)*, pages 565–571. IEEE, 2016.

[96] 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*, 44(7):3523–3542, 2021.

[97] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *Medical image analysis*, 65:101794, 2020.

[98] A. K. Mondal, J. Dolz, and C. Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*, 2018.

[99] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 7223–7226. IEEE, 2019.

[100] M. T. Nyo, F. Mebarek-Oudina, S. S. Hlaing, and N. A. Khan. Otsu's thresholding technique for mri image brain tumor segmentation. *Multimedia tools and applications*, 81(30):43837–43849, 2022.

[101] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[102] S. Pan, Y. Lei, T. Wang, J. Wynne, C.-W. Chang, J. Roper, A. B. Jani, P. Patel, J. D. Bradley, T. Liu, et al. Male pelvic multi-organ segmentation using token-based transformer vnet. *Physics in Medicine & Biology*, 67(20):205012, 2022.

[103] X. Pan, P. Luo, J. Shi, and X. Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net, Mar. 2020. arXiv:1807.09441 [cs].

[104] D. Parikesit, C. A. Mochtar, R. Umbas, and A. R. A. H. Hamid. The impact of obesity towards prostate diseases. *Prostate international*, 4(1):1–6, 2016.

[105] J. Peng and Y. Wang. Medical image segmentation with limited supervision: a review of deep network models. *IEEE Access*, 9:36827–36851, 2021.

[106] K. Preuss, N. Thach, X. Liang, M. Baine, J. Chen, C. Zhang, H. Du, H. Yu, C. Lin, M. A. Hollingsworth, et al. Using quantitative imaging for personalized medicine in pancreatic

cancer: a review of radiomics and deep learning applications. *Cancers*, 14(7):1654, 2022.

[107] Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, and D. Zhu. Attcat: Explaining transformers via attentive class activation tokens. *Advances in Neural Information Processing Systems*, 35:5052–5064, 2022.

[108] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27):e6–e6, 2021.

[109] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020.

[110] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[111] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].

[112] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.

[113] B. Schipaanboord, D. Boukerroui, D. Peressutti, J. van Soest, T. Lustberg, A. Dekker, W. van Elmpt, and M. J. Gooding. An evaluation of atlas selection methods for atlas-based automatic segmentation in radiotherapy treatment planning. *IEEE transactions on medical imaging*, 38(11):2654–2664, 2019.

[114] R. A. Schulz, J. A. Stein, and N. J. Pelc. How ct happened: the early development of medical computed tomography. *Journal of Medical Imaging*, 8(5):052110–052110, 2021.

[115] M. Segu, A. Tonioni, and F. Tombari. Batch Normalization Embeddings for Deep Domain Generalization, May 2021. arXiv:2011.12672 [cs].

[116] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE transactions on medical imaging*, 39(5):1316–1325, 2019.

[117] M. V. Sherer, D. Lin, S. Elguindi, S. Duke, L.-T. Tan, J. Cacicedo, M. Dahele, and E. F. Gillespie. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiotherapy and Oncology*, 160:185–191, 2021.

[118] L. Sun, W. Ma, X. Ding, Y. Huang, D. Liang, and J. Paisley. A 3d spatially weighted network for segmentation of brain tissue from mri. *IEEE Transactions on Medical Imaging*, 39(4):898–909, 2019.

[119] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693, 2020.

[120] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[121] M. Thor, C. Olsson, J. Deasy, D. Alsadius, N. Pettersson, A. Waldenström, G. Steineck, and J. Oh. Dose-response relationships for four gastrointestinal symptom groups in prostate cancer radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 93(3):S52, 2015.

[122] N. K. Tomar, D. Jha, U. Bagci, and S. Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 151–160. Springer, 2022.

[123] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

[124] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 547–556. IEEE, 2018.

[125] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. Multi-modal Learning from Unpaired Images: Application to Multi-organ Segmentation in CT and MRI. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 547–556, Mar. 2018.

[126] S. C. van de Leemput, M. Prokop, B. van Ginneken, and R. Manniesing. Stacked bidirectional convolutional lstms for deriving 3d non-contrast ct from spatiotemporal 4d ct. *IEEE transactions on medical imaging*, 39(4):985–996, 2019.

[127] M. Van Herk. Errors and margins in radiotherapy. In *Seminars in radiation oncology*, volume 14, pages 52–64. Elsevier, 2004.

[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*,

30, 2017.

[129] V. Venkatesh, N. Sharma, and M. Singh. Intensity inhomogeneity correction of mri images using inhomonet. *Computerized Medical Imaging and Graphics*, 84:101748, 2020.

[130] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications*, 12(1):5915, 2021.

[131] S. Wang, M. Liu, J. Lian, and D. Shen. Boundary coding representation for organ segmentation in prostate cancer radiotherapy. *IEEE transactions on medical imaging*, 40(1):310–320, 2020.

[132] T. Wang, Y. Lei, Z. Tian, X. Dong, Y. Liu, X. Jiang, W. J. Curran, T. Liu, H.-K. Shu, and X. Yang. Deep learning-based image quality improvement for low-dose computed tomography simulation in radiation therapy. *Journal of Medical Imaging*, 6(4):043504–043504, 2019.

[133] X. Wang, Z. Li, Y. Huang, and Y. Jiao. Multimodal medical image segmentation using multi-scale context-aware network. *Neurocomputing*, 486:135–146, May 2022.

[134] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[135] Z. Wu, C. Shen, and A. v. d. Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016.

[136] X. Xiao, S. Lian, Z. Luo, and S. Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018.

[137] Y. Xie, J. Zhang, C. Shen, and Y. Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[138] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[139] J. Yang, Y. Zhu, C. Wang, Z. Li, and R. Zhang. Toward unpaired multi-modal medical image segmentation via learning structured semantic consistency. In *Medical Imaging with Deep Learning*, 2023.

[140] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4956–4965, 2023.

[141] C. Yu, C. P. Anakwenze, Y. Zhao, R. M. Martin, E. B. Ludmir, J. S. Niedzielski, A. Qureshi, P. Das, E. B. Holliday, A. C. Raldow, et al. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced ct images. *Scientific reports*, 12(1):19093, 2022.

[142] O. Zavala-Romero, A. L. Breto, I. R. Xu, Y.-C. C. Chang, N. Gautney, A. Dal Pra, M. C. Abramowitz, A. Pollack, and R. Stoyanova. Segmentation of prostate and prostate zones using deep learning: A multi-mri vendor analysis. *Strahlentherapie und Onkologie*, 196:932–942, 2020.

[143] R. Zeleznik, J. Weiss, J. Taron, C. Guthier, D. S. Bitterman, C. Hancox, B. H. Kann, D. W. Kim, R. S. Punglia, J. Bredfeldt, et al. Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer. *NPJ digital medicine*, 4(1):43, 2021.

[144] G. Zhang, S. Jiang, Z. Yang, L. Gong, X. Ma, Z. Zhou, C. Bao, and Q. Liu. Automatic nodule detection for lung cancer in ct images: A review. *Computers in biology and medicine*, 103:287–300, 2018.

[145] H. Zhang, H. Li, J. R. Dillman, N. A. Parikh, and L. He. Multi-Contrast MRI Image Synthesis Using Switchable Cycle-Consistent Generative Adversarial Networks. *Diagnostics*, 12(4):816, Apr. 2022. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[146] J. Zhang, Y. Xie, Y. Xia, and C. Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021.

[147] L. Zhang, L. Lu, X. Wang, R. M. Zhu, M. Bagheri, R. M. Summers, and J. Yao. Spatio-temporal convolutional lstms for tumor growth prediction by learning 4d longitudinal patient data. *IEEE transactions on medical imaging*, 39(4):1114–1126, 2019.

[148] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021.

[149] Y. Zhang, J. Yang, J. Tian, Z. Shi, C. Zhong, Y. Zhang, and Z. He. Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 589–599. Springer, 2021.

[150] Y. Zhang, L. Zhong, H. Shu, Z. Dai, K. Zheng, Z. Chen, Q. Feng, X. Wang, and W. Yang. Cross-Task Feedback Fusion GAN for Joint MR-CT Synthesis and Segmentation of Tar-

get and Organs-At-Risk. *IEEE Transactions on Artificial Intelligence*, pages 1–12, 2022. Conference Name: IEEE Transactions on Artificial Intelligence.

[151] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018.

[152] Z. Zhang, L. Yang, and Y. Zheng. Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network, Mar. 2019. arXiv:1802.09655 [cs].

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

[154] R. Zheng, Q. Wang, S. Lv, C. Li, C. Wang, W. Chen, and H. Wang. Automatic liver tumor segmentation on dynamic contrast enhanced mri using 4d information: Deep learning model based on 3d convolution and convolutional lstm. *IEEE Transactions on Medical Imaging*, 41(10):2965–2976, 2022.

[155] Y. Zheng. Cross-modality medical image detection and segmentation by transfer learning of shapel priors. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 424–427, Apr. 2015. ISSN: 1945-8452.

[156] Y. Zhong, M. Xu, K. Liang, K. Chen, and M. Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 724–733. Springer, 2023.

[157] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[158] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 29:461–475, 2019.

[159] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[160] T. Zhou, S. Canu, P. Vera, and S. Ruan. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274, 2021.

[161] T. Zhou, S. Ruan, and S. Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019.

[162] Z. Zhou, L. Qi, X. Yang, D. Ni, and Y. Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20856–20865, 2022.

[163] Z. Zhou, L. Qi, X. Yang, D. Ni, and Y. Shi. Generalizable Cross-modality Medical Image Segmentation via Style Augmentation and Dual Normalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20824–20833, New Orleans, LA, USA, June 2022. IEEE.

[164] 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, pages 3–11. Springer, 2018.

[165] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[166] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

ABSTRACT

NOVEL TRANSFORMER ARCHITECTURES FOR 3D MULTI-MODAL AND MULTI-ORGAN MEDICAL IMAGE SEGMENTATION

by

CHENGYIN LI

June 2024

Advisor: Dr. Dongxiao Zhu

Major: Computer Science

Degree: Doctor of Philosophy

Medical image segmentation is a crucial process in medical imaging analysis, enabling precise delineation of anatomical structures and pathological regions. This dissertation explores the evolution and application of advanced deep learning models, specifically focusing on the integration of transformers and convolutional neural networks (CNNs) for enhanced medical image segmentation. The primary goal is to improve segmentation accuracy and efficiency in clinical settings, particularly for CT and MRI images.

The dissertation is structured around three key innovations. First, we introduce FocalUNETR, a novel transformer-based architecture designed to address the limitations of traditional CNNs in capturing long-range dependencies and global context in 2D CT-based prostate segmentation. FocalUNETR employs focal self-attention mechanisms and incorporates an auxiliary boundary-aware regression task to enhance segmentation precision, particularly in cases with unclear boundaries. Second, we present SwinAttUNet, a hybrid architecture combining CNNs and Swin Transformers for automatic 3D multi-organ segmentation on CT images. This approach leverages the local feature recognition capabilities of CNNs and the global contextual understanding of transformers. Third, we develop MulModSeg, a multi-modal segmentation strategy aimed at improving the segmentation of unpaired CT and MRI images. MulModSeg enhances feature extraction and model robustness by incorporating modality-conditioned text embedding and an alternating training procedure.

Extensive experiments on private and public datasets validate the effectiveness of these proposed methods. FocalUNETR achieves superior performance in 2D prostate segmentation, while SwinAttUNet outperforms state-of-the-art 3D segmentation models in both quantitative and qualitative evaluations. MulModSeg shows marked improvements in multi-modal segmentation tasks, highlighting its potential for clinical applications. This dissertation provides comprehensive frameworks for developing more accurate, efficient, and robust segmentation models, paving the way for future advancements in medical imaging and diagnostics.

AUTOBIOGRAPHICAL STATEMENT

Chengyin Li is presently a PhD candidate at Wayne State University, working under the guidance of Dr. Dongxiao Zhu in the Trustworthy AI Lab within the Department of Computer Science. He obtained his bachelor's degree from Nanjing University of Science and Technology and his master's from the University of Chinese Academy of Sciences. His research primarily focuses on medical image applications, with interests in trustworthy AI and vision-language foundation models.

Chengyin has authored several research papers on AI, which have been published in esteemed conference venues and journals such as *MICCAI, ECML, Medical Physics, and Pediatric Research*. In recognition of his outstanding contributions, he received the *Michael Conrad Award* from the Department of Computer Science at Wayne State University in 2024. He is also actively involved in research at the Henry Ford Health System, focusing on enhancing the performance and robustness of medical image segmentation tasks.