Doctoral Dissertation

ENHANCING MEDICAL IMAGE SEGMENTATION: STUDIES IN MODEL ACCURACY, PRIVACY, AND EFFICIENCY

SUN, Guanqun

Supervisor NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2024

Abstract

Medical image segmentation plays a crucial role in quantifying diseases, assessing prognosis, and evaluating treatment results. However, manual segmentation is time-consuming, prone to interobserver variability, and limited by the availability of skilled experts. Despite advances in deep learning-based approaches for automatic segmentation, challenges such as precise boundary delineation, limited annotated data, and the trade-off between model complexity and performance persist. In addition, data privacy concerns hinder the sharing of medical images between institutions, preventing collaborative research and model development.

This dissertation focuses on improving medical image segmentation by addressing three key aspects: model accuracy, data privacy, and computational efficiency. This dissertation proposes novel deep learning architectures and techniques that leverage the power of attention mechanisms, transformer models, federated learning, and knowledge distillation to tackle these challenges.

Firstly, this dissertation introduce DA-TransUNet, a dual attention transformer U-Net architecture that integrates spatial and channel attention mechanisms with transformer models. DA-TransUNet effectively captures fine-grained details and long-range dependencies in medical images, leading to improved segmentation accuracy compared to state-of-the-art methods.

Secondly, this dissertation proposes MIPC-Net, a mutual inclusion mechanism for precise boundary segmentation. MIPC-Net uses complementary information from position and channel features to enhance the delineation of complex anatomical structures and small lesions, resulting in a more accurate boundary segmentation.

Thirdly, this thesis introduces FKD-Med, a framework for medical image segmentation that prioritizes privacy and optimizes communication. FKD-Med integrates federated learning and knowledge distillation techniques to enable collaborative model training between multiple institutions while preserving data privacy. It also improves model efficiency by distilling knowledge from complex models to lighter ones, reducing computational requirements without compromising segmentation performance.

Extensive experiments on multiple benchmark datasets demonstrate the superior performance of the proposed methods and frameworks in terms of segmentation accuracy, boundary precision, and computational efficiency. The contributions of this dissertation advance the field of medical image

segmentation by proposing novel architectures, mechanisms, and frameworks that address key challenges related to model accuracy, data privacy, and computational efficiency.

Future research directions include exploring additional attention mechanisms and transformer variants, extending the proposed methods to 3D and volumetric segmentation tasks, integrating differential privacy techniques for enhanced data protection, developing advanced model compression and acceleration techniques, investigating the generalizability and transferability of the proposed approaches to different medical imaging modalities and anatomical regions, and improving the interpretability and explainability of the segmentation models.

By advancing the state-of-the-art in medical image segmentation, this dissertation contributes to the development of accurate, privacy-preserving, and efficient segmentation models that can be seamlessly integrated into clinical workflows, ultimately improving patient care through more precise diagnosis, treatment planning, and monitoring of various diseases and conditions.

Keywords: Medical Image Segmentation, Dual Attention, Mutual Inclusion, Federated Learning, Knowledge Distillation.

Acknowledgment

I extend my heartfelt gratitude to all those who have supported me throughout my Ph.D. journey. First, I am deeply indebted to my primary supervisor, Prof. NGUYEN Le-Minh, for his unwavering guidance, expertise, and mentorship. His dedication and commitment to my research have been instrumental in shaping my work and helping me grow as a scholar. From guiding me through the development of my major project proposal to providing detailed feedback on my journal papers and dissertation outline, his support has been invaluable.

I am also immensely grateful to my second supervisor, Prof.INOUE Naoya, whose contributions have been equally significant. His guidance on my proposal and dissertation outline helped me refine my ideas and present them with clarity and coherence. His perspective and input have enriched my work and broadened my understanding of the field. I express my sincere appreciation to my minor project advisor, Dr.RACHARAK, Teeradaj, whose valuable insights and feedback have contributed significantly to my research and the publication of several journal articles.

I would like to extend my deepest thanks to the examiners of my dissertation: OKADA Shogo, HASEGAWA Shinobu, SHIRAI Kiyoaki, and MA Jianhua. Your professional suggestions and meticulous review have greatly improved the quality of this dissertation. I am sincerely grateful for your valuable feedback and support.

To my fellow students and colleagues at the Japan Advanced Institute of Science and Technology, I am grateful for the stimulating discussions, shared experiences, and the sense of camaraderie that have made this journey both rewarding and enjoyable. Your support and friendship have been invaluable.

I also thank the Japan Advanced Institute of Science and Technology for providing an exceptional academic environment, resources, and opportunities that have facilitated my research and professional development.

To my colleagues at Hangzhou Medical College, including ZHU Xiaofeng, YANG Tianhua, HU Haixiang, TAO Ying, YE Hanfeng, DONG Jingjing, LIU Yaru, MA Lin, ZHENG Leyi, XI Ningli, LIAO Zhener, and LIU Zhi, I extend my appreciation for their understanding, flexibility, and support as I pursued my doctorate degree while working alongside them. I am particularly grateful to Prof.XIN Junyi for his unwavering support and tireless advocacy on my behalf. His dedication to ensuring that I could successfully balance my work and doctoral studies has been a cornerstone of my journey. Through

his invaluable guidance, encouragement, and belief in my potential, he has played a crucial role in my ability to navigate the challenges of pursuing my professional and academic aspirations. His commitment to my success has had a profound impact on my life and for that I am deeply thankful.

Last but not least, I would like to thank my coauthors, Feihe Shao, Weikun Kong, Zichang Xu, Jianhua Ma, Xin Liu, Jianan Chen, and Qiang Ma, for their collaboration and contributions to our joint publications. Their expertise and insights have enriched my research and expanded my understanding of the field.

Finally, I would like to express my deepest gratitude to my parents for their unwavering support throughout my academic journey. Their constant encouragement and belief in me have been the driving force behind my ability to pursue multiple degrees and ultimately undertake my Ph.D. studies. Without their love, sacrifices, and dedication to my education, I would not have had the opportunity to explore my passions and reach this significant milestone. I am forever grateful for the strong foundation they have provided me, which has been instrumental in shaping my academic and personal growth.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Medical image segmentation is a fundamental task in medical image analysis, playing a crucial role in quantifying diseases, assessing prognosis, and evaluating treatment outcomes. It involves delineating regions of interest within medical images, such as organs, lesions, or other anatomical structures. Accurate and efficient segmentation is essential for clinical decision-making and patient care.

Traditionally, medical image segmentation has been performed manually by skilled professionals, such as radiologists and clinical experts. However, manual segmentation is a time-consuming and labor-intensive process, often taking hours or even days to complete for a single patient. Moreover, manual segmentation is prone to variability between observers, as different experts may have different interpretations and delimitations of the same image, leading to inconsistencies in the results [1].

With rapid advances in deep learning technologies, automatic medical image segmentation has gained significant attention in recent years. Deep learning-based approaches, such as convolutional neural networks (CNNs) and their variants, have shown promising results in segmenting various anatomical structures and lesions across different imaging modalities. These approaches aim to improve the efficiency and precision of the segmentation process, alleviating the burden on medical professionals and enabling more consistent and reproducible results.

Medical image segmentation differs from general segmentation tasks in several key aspects. First, target regions in medical images often exhibit irregular shapes, ambiguous boundaries, and complex anatomical structures, while general segmentation tasks typically involve objects with more regular shapes and clearer boundaries. Second, medical images are heterogeneous, originating from various imaging modalities (e.g. CT, MRI, X-ray) and presenting anatomical variations among patients. In contrast, general segmentation tasks often deal with images from the same domain,

which feature relatively consistent characteristics. Third, medical image datasets are usually smaller and more costly to annotate compared to large-scale annotated datasets available for general segmentation tasks. Finally, medical image segmentation requires high precision, especially along the boundaries of target regions, as accuracy directly impacts diagnosis and treatment decisions. General segmentation tasks, on the other hand, may have lower accuracy requirements for object boundaries. These distinct characteristics of medical image segmentation require the development of specialized algorithms and techniques that can effectively address the unique challenges posed by medical images.

Despite the progress made in medical image segmentation using deep learning, several challenges remain. One major challenge is the need for precise boundary delineation, especially for complex anatomical structures and small lesions. Accurate boundary segmentation is crucial for treatment planning and surgical interventions, where even minor inaccuracies can have significant clinical consequences. Another challenge is the limited availability of annotated data, as manual annotation of medical images is a time-consuming and expensive process. This scarcity of labeled data hinders the development and generalization of deep learning models. Furthermore, there is often a trade-off between model complexity and performance, as more complex models may achieve higher segmentation accuracy, but at the cost of increased computational requirements and longer inference times [1–3].

In addition to these challenges, data privacy and security concerns are of the utmost importance in the medical domain. Medical images contain sensitive patient information, and sharing such data between different institutions for collaborative research and model development is often restricted by privacy regulations and ethical considerations. This poses a significant barrier to the use of large-scale datasets and the use of collective knowledge from multiple institutions to improve segmentation models.

Motivated by these challenges and the potential impact of accurate and efficient medical image segmentation on patient care, this dissertation aims to explore novel techniques and frameworks to enhance the performance, privacy, and efficiency of medical image segmentation using deep learning.

1.2 Research Objectives

The main objectives of this research are as follows:

1. To develop advanced deep learning architectures that can effectively capture fine-grained details and long-range dependencies in medical images for accurate segmentation.

2. Investigate techniques for precise boundary delineation of complex anatomical structures and small lesions.
3. Explore privacy-preserving approaches for collaborative model training across multiple institutions while ensuring data security and confidentiality.
4. Improve the computational efficiency of medical image segmentation models, enabling faster inference and deployment in clinical settings.

1.3 Research Framework

To better illustrate the research objectives and the relationships between the key components of this dissertation, a research framework is presented in Figure 1.1. The research framework, as shown in Figure 1.1, revolves

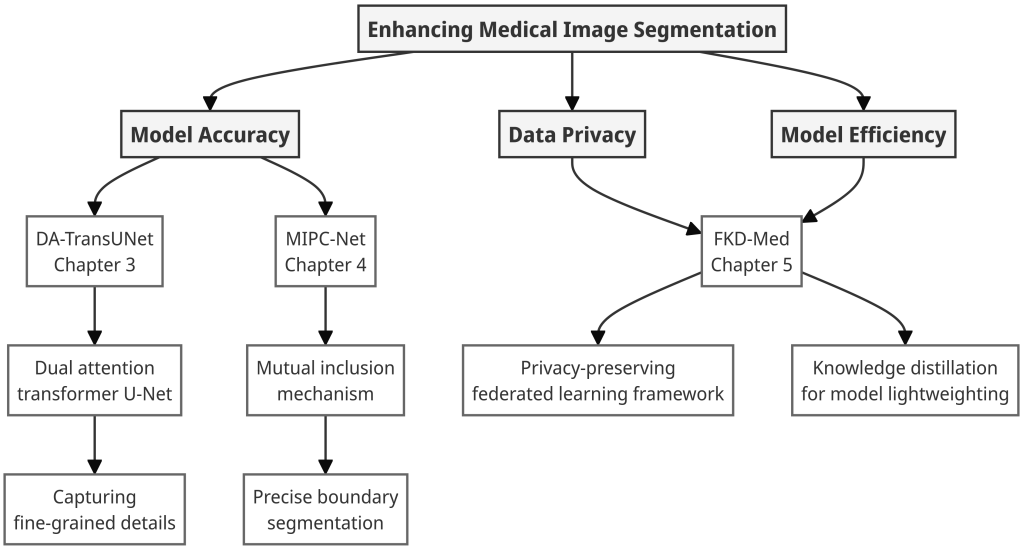

Figure 1.1: Research framework of the dissertation, highlighting the main research directions and their corresponding chapters.

around the central goal of enhancing medical image segmentation. This goal is approached from three main research directions: model accuracy, data privacy, and model efficiency. In the direction of model accuracy, two novel architectures are proposed: DA-TransUNet (Chapter 3) and MIPC-Net (Chapter 4). DA-TransUNet integrates dual attention mechanisms and transformer models into a U-Net architecture to capture fine-grained details in medical images. MIPC-Net introduces a mutual inclusion mechanism for

precise boundary segmentation, leveraging complementary information from position and channel features. The data privacy direction is addressed by FKD-Med (Chapter 5), a federated learning framework that preserves privacy that allows the training of collaborative models across multiple institutions while ensuring data confidentiality. FKD-Med also contributes to the model efficiency direction by employing knowledge distillation techniques for model lightweighting, reducing computational requirements without compromising segmentation performance. This research framework provides a clear overview of the main contributions and their interconnections, guiding the structure and content of the dissertation.

1.4 Contributions

The main contributions of this dissertation are summarized as follows:

- This dissertationpropose DA-TransUNet, a novel dual attention transformer U-Net architecture that integrates spatial and channel attention mechanisms with transformer models (Chapter 3). DA-TransUNet effectively captures fine-grained details and long-range dependencies in medical images, leading to improved segmentation accuracy compared to existing methods.
- This dissertation introduces MIPC-Net, a mutual inclusion mechanism for precise boundary segmentation (Chapter 4). MIPC-Net uses complementary information from position and channel features to enhance the delineation of complex anatomical structures and small lesions, resulting in a more accurate boundary segmentation.
- This dissertation presents FKD-Med, a privacy-sensitive communication-optimized framework for medical image segmentation (Chapter 5). FKD-Med integrates federated learning and knowledge distillation techniques to enable collaborative model training across multiple institutions while preserving data privacy. It also improves model efficiency by distilling knowledge from complex models to lighter ones, reducing computational requirements without compromising segmentation performance.
- This dissertation conducts extensive experiments on multiple benchmark datasets to evaluate the effectiveness of the proposed methods and frameworks. The results demonstrate the superior performance of our approaches compared to state-of-the-art methods in terms of segmentation accuracy, boundary precision, and computational efficiency.

1.5 Thesis Organization

The remainder of this dissertation is organized as follows:

- Chapter 2 provides a comprehensive literature review of the recent advances in medical image segmentation. Covers various topics, including U-Net and its variants, attention mechanisms, transformer-based models, federated learning, and knowledge distillation. The limitations of existing approaches and potential research directions are also discussed.
- Chapter 3 introduces DA-TransUNet, a dual attention transformer U-Net architecture for medical image segmentation. The architecture design, integration of attention mechanisms and transformer models, and experimental results are presented in detail.
- Chapter 4 presents MIPC-Net, a mutual inclusion mechanism for precise boundary segmentation. The methodology, including the extraction and fusion of position and channel features, is described, along with the experimental evaluation and comparison with state-of-the-art methods.
- Chapter 5 describes FKD-Med, a privacy-sensitive privacy-optimized communication framework for medical image segmentation. The integration of federated learning and knowledge distillation techniques, as well as the experimental setup and results, is discussed in depth.
- Chapter 6 concludes the dissertation by summarizing the main findings, contributions, and potential future research directions. It also highlights the impact of the proposed methods and frameworks on the advancement of medical image segmentation and improving patient care.

Chapter 2

Literature Review

2.1 Overview of Medical Image Segmentation Techniques

Medical image segmentation plays a crucial role in quantifying diseases, assessing prognosis, and evaluating treatment results. It involves delineating regions of interest within medical images, such as organs, lesions, or other anatomical structures. However, manual segmentation by experienced professionals is time-consuming and prone to variability between observers [1]. With the advent of deep learning technologies, automatic medical image segmentation has gained significant attention in the research community, with the aim of improving the efficiency and accuracy of the segmentation process.

Despite the progress made in medical image segmentation, several challenges remain. These include the need for precise boundary delineation, limited availability of annotated data, and the trade-off between model complexity and performance [1–3]. In addition, medical image segmentation differs from generic image segmentation tasks, as it requires capturing fine-grained details and handling variations in anatomy and pathology between patients.

This chapter provides a comprehensive overview of recent advances in medical image segmentation, focusing on deep learning-based approaches. The discussion begins with the U-Net architecture and its variants, which have been widely adopted in medical image segmentation. The application of attention mechanisms, particularly models based on dual attention and transformers, is explored to enhance segmentation performance. The challenges of data privacy and scarcity are discussed, and the discussion focuses on how federated learning can address these issues. Finally, the role of knowledge distillation in improving model efficiency and performance is examined. Throughout the chapter, the limitations of existing approaches are highlighted and potential research directions to advance the field of medical image segmentation are identified. By addressing these challenges and

exploring innovative solutions, the aim is to contribute to the advancement of medical image segmentation techniques and ultimately improve patient care through more accurate and efficient disease quantification, prognosis evaluation, and treatment evaluation.

2.2 U-Net and Its Variants in Medical Image Segmentation

2.2.1 U-Net Architecture

U-Net is a widely adopted architecture in medical image segmentation, known for its efficient use of data augmentation and its ability to achieve superior performance even with limited datasets [4].

2.2.2 Variants of U-Net

Building upon the U-Net architecture, various variants have been proposed to further enhance its segmentation performance. ResUNet [5] incorporates residual connections to improve segmentation, particularly in the context of polyp detection during colonoscopy examinations. Attention U-Net [6] integrates attention mechanisms to boost the localization and segmentation of the pancreas. Other notable variants include DAResUNet [7], which combines double attention and residual mechanisms, and Attention Res-UNet [8], which explores the substitution of hard attention with soft attention.

TransUNet [9] represents a significant advancement by innovatively combining the Transformer architecture with the U-Net structure. Subsequent works, such as TransU-Net++ [10], build on this foundation by incorporating attention mechanisms into skip connections and feature extraction. Swin-Unet [11] goes one step further by replacing every convolution block in U-Net with the Swin-Transformer [12]. DS-TransUNet [13] proposes the integration of a multiscale Transformer module (TIF) into skip connections, while AA-TransUNet [14] leverages the Block Attention Model (CBAM) and Deep Separable Convolution (DSC) to optimize TransUNet.

2.2.3 Skip Connections and Model Integration

Skip connections play a vital role in U-Net-based models, aiming to bridge the semantic gap between the encoder and decoder and effectively recover fine-grained object details [15] [16] [17]. Modifications to skip connections can be categorized into three primary approaches. The first focuses on increasing

the complexity of skip connections, as exemplified by the Dense-like structure in U-Net++ [18] and the full-scale skip connections in U-Net3++ [19]. The second approach involves processing feature maps within skip connections, such as the 3D hybrid residual attention-aware method introduced in RA-UNet [20] for precise feature extraction. The third approach combines feature maps of encoder and decoder, as seen in BCDU-Net [21], where a bidirectional convolutional long short-term memory (LSTM) module is added to the skip connections.

Model integration techniques have also been explored to enhance the learning capacity of deep neural networks. DAResUNet [7] incorporates residual modules and dual attention blocks into skip connections, while DS-TransUNet [13] merges Transformer mechanisms into skip connections. IB-TransUNet [3] integrates a multiresolution fusion mechanism into skip connections, and DA-TransUNet [2] optimizes skip connections using image feature positions and channels. However, these approaches often focus on specific aspects of the model, lacking a comprehensive consideration of the overall structure.

2.3 Attention Mechanisms in Medical Image Segmentation

2.3.1 Overview of Attention Mechanisms

Attention mechanisms have gained significant traction in medical image segmentation due to their ability to guide the model's focus towards relevant features and improve performance. The concept of attention mechanisms was first introduced in the context of machine translation [22] and has since been applied to various domains, including image generation [23], image captioning [24], and visual question answering [25].

2.3.2 Channel Attention and Spatial Attention

Two primary types of attention mechanisms have been explored in medical image segmentation: channel attention and spatial attention. Channel attention focuses on assigning importance to different channels of the feature maps, while spatial attention emphasizes the importance of different spatial locations. The Squeeze-and-Excitation (SE) block [26] is a popular choice for channel attention, while the Convolutional Block Attention Module (CBAM) [27] incorporates both channel and spatial attention. These attention mech-

anisms have been successfully integrated into various U-Net-based models, such as Attention U-Net [6] and SA-UNet [28].

2.3.3 Dual Attention Mechanisms

Dual attention mechanisms, which combine channel attention and spatial attention, have demonstrated promising results in medical image segmentation. The Dual Attention Network (DANet) [29] employs position and channel attention modules to capture long-range dependencies and improve segmentation accuracy. The Multilevel Dual Attention U-Net [30] integrates dual attention modules at different scales to enhance polyp segmentation.

2.4 Transformer-based Models in Medical Image Segmentation

2.4.1 Overview of Transformer Models

Transformer models, initially proposed for natural language processing tasks [31], have recently gained traction in computer vision, including medical image segmentation. The self-attention mechanism in Transformers allows for capturing long-range dependencies and global context, which is particularly beneficial for medical images with complex structures and variations.

2.4.2 Transformer-based Architectures for Medical Image Segmentation

TransUNet [9] pioneered the application of Transformers in medical image segmentation by incorporating Transformer layers into the encoder of the U-Net architecture. Swin-Unet [11] further advanced this approach by introducing a pure transformer-based U-shaped encoder-decoder architecture, using the power of the Swin-Transformer [12]. Other notable Transformer-based architectures include DS-TransUNet [13], which integrates a multiscale Transformer module (TIF) into skip connections, and MT-UNet [32], which incorporates a Mixed Transformer module for enhanced feature extraction.

2.4.3 Integration of Transformer and CNN

Efforts have also been made to combine the strengths of Transformers and convolutional neural networks (CNNs) for medical image segmentation. TransFuse [33] introduces a fusion of CNNs and Transformers, utilizing

the BiFusion module and the AG block to combine features from both architectures. ResViT [34] integrates the contextual sensitivity of vision transformers, the precision of convolution operators, and the realism of adversarial learning. These hybrid approaches aim to leverage the global context captured by Transformers and the local details extracted by CNNs to improve segmentation accuracy.

2.5 Federated Learning in Medical Image Segmentation

2.5.1 Overview of Federated Learning

Federated learning (FL) has emerged as a promising approach to address data privacy and scarcity issues in medical image segmentation. FL enables collaborative model training across multiple institutions without the need for direct data sharing, thus preserving patient privacy and adhering to regulatory requirements.

2.5.2 Applications of Federated Learning in Medical Image Segmentation

Several studies have explored the application of FL in medical image segmentation. [35] introduced a distributed real-time network framework and provided a comprehensive analysis of different FL methods for the segmentation of brain tumors. [36] investigated the integration of differential privacy techniques to strike a balance between model performance and privacy protection. [37] proposed a scalable FL framework with a U-Net architecture, achieving significant improvements in brain tumor segmentation while ensuring advanced data privacy and security measures.

Other notable contributions include FedMix [38], which addresses the varying levels of image supervision across local clients by dynamically adjusting the aggregation weights, and FedSeg [39], which tackles the challenges of non-IID data distribution and class heterogeneity in FL for semantic segmentation.

2.5.3 Challenges and Future Directions

Despite the progress made in applying FL to medical image segmentation, several challenges remain. These include the communication overhead associated with transferring model updates between clients and the server,

the potential for model performance degradation due to data heterogeneity across clients, and the need for effective aggregation strategies to handle non-IID data distributions. Future research directions may explore the integration of FL with other techniques, such as knowledge distillation and model compression, to further improve the efficiency and performance of FL in medical image segmentation.

2.6 Knowledge Distillation in Medical Image Segmentation

2.6.1 Overview of Knowledge Distillation

Knowledge distillation (KD) is a technique that aims to transfer knowledge from a large, complex teacher model to a smaller, more efficient student model. Using the knowledge learned by the teacher model, KD enables the student model to achieve comparable performance with reduced computational complexity and memory requirements.

2.6.2 Applications of Knowledge Distillation in Medical Image Segmentation

KD has been applied to various tasks in medical image segmentation to improve the efficiency and performance of the model. The adaptive perspective distillation approach (APD) [40] introduces an adaptive local perspective for each training sample, enhancing the KD process. Cross-Image Relational Knowledge Distillation (CIRKD) [41] focuses on transferring structured relations, such as pixel-to-pixel and pixel-to-region correlations, between images.

In the context of medical imaging, [42] proposed an efficient architecture that combines improved segmentation capability with runtime efficiency. [43] introduced a novel methodology that integrates two individual segments, each focusing on obtaining modality-specific knowledge. The Structural and Statistical Texture Knowledge Distillation (SSTKD) framework [44] leverages both structural and statistical texture knowledge to enhance the KD process.

2.6.3 Challenges and Future Directions

The integration of KD with FL for medical image segmentation remains largely unexplored. Although some studies have proposed frameworks that

combine FL and KD, such as FedDKD [45], FedICT [46], and MetaFed [47], more research is needed to understand how to effectively integrate these techniques in the context of medical image segmentation. Investigating the synergies between KD and FL, and developing efficient and privacy-preserving KD strategies represent promising avenues for future research.

2.7 Chapter Summary

In this chapter, this dissertation has provided a comprehensive overview of the recent advances in medical image segmentation, focusing on deep learning-based approaches. This dissertation discussed the U-Net architecture and its variants, which have been widely adopted in the field, and highlighted the importance of skip connections and model integration techniques. This dissertation also explored the application of attention mechanisms, particularly dual attention- and transformer-based models, to capture long-range dependencies and improve segmentation performance.

Furthermore, this dissertation addressed the challenges of data privacy and scarcity in medical image segmentation and discussed how federated learning can enable collaborative model training across multiple institutions while preserving patient privacy. This dissertation also examined the role of knowledge distillation in improving model efficiency and performance and identified the potential for integrating knowledge distillation with federated learning.

Throughout the chapter, this dissertation emphasized the limitations of existing approaches and identified potential research directions to advance the field of medical image segmentation. These include the development of more comprehensive model integration strategies, the exploration of novel attention mechanisms tailored to medical images, the investigation of efficient and privacy-preserving federated learning techniques, and the integration of knowledge distillation with federated learning.

By addressing these challenges and exploring innovative solutions, the aim is to contribute to the advancement of medical image segmentation techniques and ultimately improve patient care through more accurate and efficient disease quantification, prognosis evaluation, and treatment evaluation.

Chapter 3

DA-TransUNet: Dual Attention Transformer U-Net for Accurate Medical Image Segmentation

3.1 Motivation and Objectives

Machine learning and deep learning techniques have emerged as powerful tools in biomedical research, revolutionizing disease diagnosis, treatment planning, and personalized medicine [48, 49]. Medical image segmentation is the process of delineating regions of interest within medical images for diagnosis and treatment planning. It serves as a cornerstone in medical image analysis. Manual segmentation is accurate and affordable for pathology diagnosis, but is vital in standardized clinical settings. In contrast, automated segmentation ensures a reliable and consistent process, enhancing efficiency, cutting down on labor and costs, and preserving accuracy. Consequently, there is a substantial demand for exceptionally accurate automated medical image segmentation technology within the realm of clinical diagnostics. However, medical image segmentation faces unique challenges, such as the need for precise delineation of complex anatomical structures, variability between patients, and the presence of noise and artifacts in images [50]. These challenges require the development of advanced segmentation techniques that can capture fine-grained details while maintaining robustness and efficiency.

In the past decade, the U-Net architecture has emerged as the cornerstone of various segmentation tasks, consistently delivering impressive results. The original U-Net model [51], along with its subsequent enhancements, has achieved remarkable success. Notable variants have emerged during this period, such as ResUnet [5], which incorporates residual learning concepts, and UNet++ [18], which focuses on optimizing skip connections. While these CNN-based approaches have dominated the field, the introduction of the Transformer architecture has ushered in a paradigm shift. Originally conceived for sequence-to-sequence modeling in Natural Language Processing

(NLP) [52], Transformers have since found significant applications in Computer Vision (CV). Vision Transformers (ViTs) [53], for instance, segment images into patches and process their embeddings through a transformer network, achieving strong performance. This development marks a significant trend towards more flexible and powerful models, moving beyond traditional CNN architectures. The shift from CNNs to Transformers represents a fundamental change in approach, offering new possibilities to improve segmentation tasks in medical imaging. Although the above-mentioned U-Net structures have enhanced the capabilities of models in segmentation tasks [51] [5] [18], they do not integrate the more powerful feature extraction abilities inherent in the Transformer and attention mechanisms, which limits their potential for further improvement. On the one hand, several studies have made progress in image segmentation by leveraging Dual Attention (DA) mechanisms for both channels and positions. The Dual Attention Network (DANet) uses a Position Attention Block (PAM) and a Channel Attention Block (CAM) from the DA Network to segment images of natural scenes [54]. This research focuses primarily on scene segmentation and does not explore the unique characteristics of medical imagery. In addition, DAResUnet [7] introduces a dual attention block combined with a residual block (Res-Block) in a U-net architecture for medical image segmentation, demonstrating significant improvements in this domain. However, in the realm of medical image segmentation, existing models, including those employing Dual Attention mechanisms, have not yet extensively explored the optimal integration of Dual Attention with Transformer models for enhanced feature extraction; this oversight represents a significant research opportunity in the task of medical image segmentation. Therefore, addressing this gap and optimizing the integration of Transformers and Dual Attention mechanisms in the context of medical image segmentation poses a significant challenge for future research in the field.

To overcome the above drawbacks, recent studies have explored the application of Transformer models in medical image segmentation. Inspired by ViTs, TransUNet [9] further combines the functionality of ViTs with the advantages of U-net in the field of medical image segmentation. Specifically, it employs a transformer's encoder to process the image and employs CNN and hopping connections for accurate up-sampling feature recovery, yet it neglects image-specific features like position and channel. These aspects are crucial in capturing the nuanced variations and complex structures that often present in medical images, which are essential for accurate diagnosis and analysis. Swin-Unet [11] combines the Swin transform block with the U-net structure and achieves good results. However, adding extensive Transformer blocks inflates the parameter count without significantly improving

results. This study merely stacked multiple Transformers to enhance models, resulting in inflated parameters and computational complexity with marginal gains in performance. In addition, some studies have specifically focused on incorporating position and channel attention mechanisms in medical image segmentation. For example, DA-DSUnet has been applied to head and neck tumor segmentation, but it does not combine the position attention module (PAM) and the channel attention module (CAM), nor does it discuss the potential filtering role of DA blocks in skip connections [55]. Additionally, it does not leverage ViT for feature extraction. Another example is research on brain tumor segmentation, which, when applying DA blocks, limits its scope to brain tumors without validating other types of medical images [56]. These studies integrate DA blocks with other blocks but do not thoroughly explore the role of DA in skip connections or optimize DA blocks for the unique intricacies of medical imaging.

However, despite the progress made by these transformer-based approaches, they often overlook the importance of integrating image-specific features, such as position and channel information, which are crucial for capturing the nuanced variations and complex structures in medical images. In addition, existing methods that incorporate dual attention mechanisms have not been optimized for the unique characteristics of medical images, leaving room for further improvement. To address these limitations, this dissertation proposes DA-TransUNet, which strategically integrates the Dual Attention Block (DA-Block) into the transformer-based U-Net architecture, specifically tailored for medical image segmentation.

In this chapter, the proposed model DA-TransUNet is an innovative approach to medical image segmentation that integrates the Transformer mechanism, specifically the Vision Transformer (ViT) and a Dual Attention (DA) mechanism within a U-Net architecture. First, the ViT Transformer is combined with DA in the U-Net structure encoder, enhancing feature extraction capabilities by leveraging the detailed characteristics of medical images. This integration allows the model to capture both local and global contextual information, which is essential for accurate segmentation of complex anatomical structures. Then, to further refine the feature extraction tailored to medical images, DA is optimized for specific channels and incorporated into every module of the skip connections, allowing the model to effectively filter out irrelevant information and focus on the most discriminative features. Skip connections pass the shallow positional information from the encoder, while the DA module refines the crucial detailed features. This targeted optimization is substantiated by extensive ablation studies, demonstrating its significance in improving the model's performance. Lastly, this architecture has been rigorously tested in five medical image segmentation datasets and

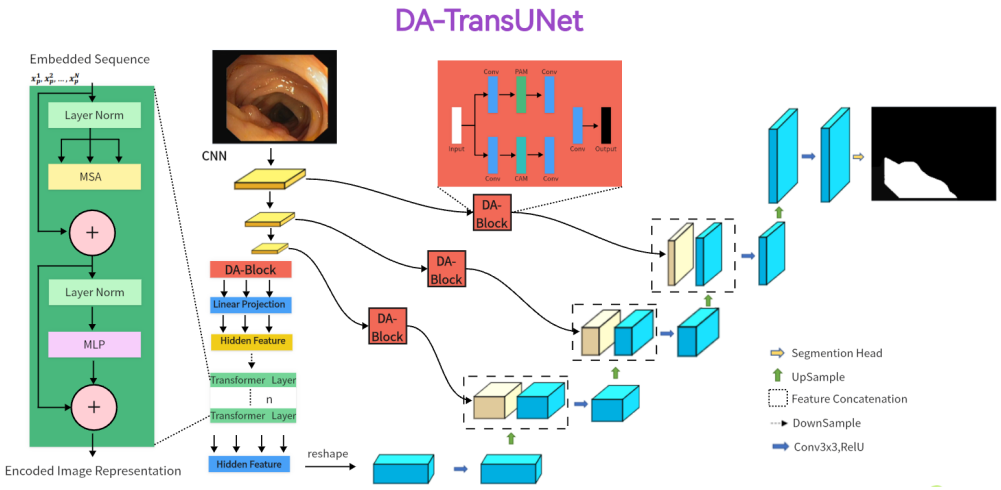

Figure 3.1: Illustration of the proposed dual attention transformer U-Net(DA-TransUNet).

extensive ablation studies, demonstrating its effectiveness and superiority [57] [58] [59, 60] [61] [62] [63, 64].

The main contributions of this chapter are summarized as follows:

1) The model of DA-TransUnet is proposed by integrating Transformer ViT and Dual Attention in U-net architecture's encoder and skip connections. This design enhances feature extraction capabilities to better extract detailed features of medical images.
2) This dissertation proposes an optimized dual attention (DA) block that is designed for medical image segmentation with two key enhancements: the optimization of intermediate channel configurations within the DA block, and its integration into each skip-connection layer to effectively filter irrelevant information. These are validated through comprehensive ablation experiments.
3) The segmentation performance and generalizability of DA-TransUnet are validated on five medical datasets. Compared to recent related studies, DA-TransUnet exhibits superior results in medical image segmentation, demonstrating its effectiveness in this field.

3.2 Methodology

Before delving into the details of our proposed DA-TransUNet architecture, it is essential to understand the concept of channels in digital images and

convolutional neural networks (CNNs).

In digital images, a channel refers to a specific component of the image data. For grayscale images, there is only one channel that represents the intensity of each pixel. However, for color images, there are typically three or four channels, each representing a different color component. The most common color space is RGB (Red, Green, Blue), where each pixel is described by the intensity of these three color channels. In some cases, a fourth channel, known as the alpha channel, is used to represent transparency. The number of channels and their interpretation depend on the specific image format and color space being used.

In the context of convolutional neural networks (CNNs), the concept of channels extends beyond just the input image. As the data flows through the layers of the network, each layer typically has multiple channels, also known as feature maps. These channels can be thought of as different "views" or "aspects" of the input data, each learning to detect specific features or patterns. The number of channels in a layer is a hyperparameter that can be adjusted to control the complexity and capacity of the network.

In the subsequent section,this study proposes the DA-TransUNet architecture, illustrated in Figure.3.1. This dissertation starts with a comprehensive overview of the architecture. Next,this study detailed the architecture's key components in the following order: the dual attention blocks(DA-Block), the encoder, the skip connections, and the decoder.

3.2.1 Overview of DA-TransUNet

Figure 3.1 illustrates the innovative architecture of DA-TransUNet, which is composed of three fundamental components: an encoder, a decoder, and skip connections. This design represents a significant advancement over traditional segmentation models.

The encoder in DA-TransUNet is distinguished by its hybrid structure, which seamlessly integrates a conventional convolutional neural network (CNN) with a transformer layer. This fusion is further enhanced by the novel Dual Attention Block (DA-Block), a key innovation exclusive to this architecture. In contrast, the decoder maintains a more traditional structure, primarily utilizing conventional convolutional mechanisms for upsampling and feature reconstruction.

A crucial aspect of DA-TransUNet's design is the optimization of skip connections through the strategic placement of DA-Blocks. These blocks serve a dual purpose: they act as information filters in skip connections, effectively reducing noise and irrelevant data, while simultaneously enhancing the accuracy of image reconstruction by preserving and emphasizing salient

features. This approach marks a departure from both traditional convolutional methodologies and models that rely heavily on transformer architectures. DA-TransUNet's unique use of DA-Blocks enables the extraction and utilization of image-specific positional and channel features, significantly boosting the model's overall performance.

When compared to conventional U-Net architectures, DA-TransUNet offers several advantages. The incorporation of a transformer layer in the encoder facilitates the capture of global dependencies, a capability lacking in the purely convolutional approach of U-Net. Additionally, the inclusion of DA-Blocks in both the encoder and skip connections enhances the model's ability to extract and utilize image-specific positional and channel features. This combination allows for more effective capture of fine-grained details, which is crucial in medical image segmentation tasks.

The rationale behind DA-TransUNet's design stems from a critical analysis of the strengths and limitations of both U-Net architectures and transformers in feature extraction. Transformers excel in global feature extraction through self-attention mechanisms but are limited by their unidirectional focus on positional attributes. Conversely, traditional U-Net architectures are adept at local feature extraction but lack comprehensive global contextualization capabilities.

To address these constraints, DA-TransUNet strategically integrates DA-Blocks both before the transformer layers and within the encoder-decoder skip connections. This design achieves two primary objectives: it refines the feature map input to the transformer, enabling more nuanced and precise global feature extraction, and it optimizes the features transmitted through skip connections, facilitating more accurate feature map reconstruction in the decoder.

In conclusion, DA-TransUNet's architecture successfully combines the strengths of both U-Net and transformer-based models while mitigating their respective weaknesses. The result is a robust system capable of advanced, image-specific feature extraction, particularly suited for the complexities of medical image segmentation.

3.2.2 Dual Attention Block(DA-Block)

The Dual Attention Block (DA-Block), depicted in Figure 3.2, represents a novel approach to feature extraction in image segmentation tasks. This innovative module is designed to capture and integrate both positional and channel-specific information, allowing for a more comprehensive representation of image characteristics.

In the context of U-Net-style architectures, the DA-Block's specialized

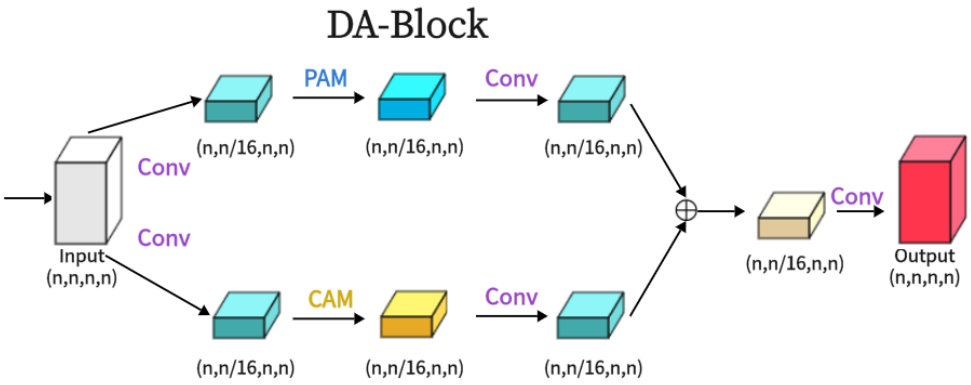

Figure 3.2: The proposed Dual Attention Block (DA-Block) is shown in the Figure.

feature extraction capabilities play a pivotal role. While traditional Transformer models excel at leveraging attention mechanisms for global feature extraction, they often lack the specificity required for capturing image-unique attributes. The DA-Block addresses this limitation by excelling in both position-based and channel-based feature extraction, resulting in a more nuanced and accurate feature set.

The integration of the DA-Block into both the encoder and skip connections of our model significantly enhances its segmentation performance. This strategic placement allows for refined feature propagation throughout the network, contributing to improved overall accuracy.

At its core, the DA-Block comprises two essential components: a Position Attention Module (PAM) and a Channel Attention Module (CAM). The PAM focuses on capturing spatial relationships within the image, while the CAM emphasizes the importance of channel-wise information. These modules are adapted from the Dual Attention Network, originally proposed for scene segmentation tasks [54]. By repurposing these components for medical image analysis, our model achieves a more thorough and context-aware feature extraction process.

The synergy between PAM and CAM within the DA-Block enables our model to simultaneously consider spatial configurations and feature channel correlations. This dual-focus approach results in a richer, more comprehensive representation of the input image, particularly beneficial for the intricate task of medical image segmentation where both local details and global context are crucial.

PAM (Position Attention Module): As illustrated in Figure 3.3, the Position Attention Module (PAM) is designed to capture and leverage spatial dependencies between different positions on feature maps. This module operates by updating specific features through a weighted sum of all positional features, where the weights are determined by the similarity of features between any two given positions. This mechanism enables PAM to effectively extract meaningful spatial features.

The process begins with PAM taking a local feature A $\in R^{C\times H\times W}$ as input, where C represents the number of channels, and H and W denote the height and width, respectively. This input is then fed through a convolutional layer, producing three new feature maps: B, C, and D, each with dimensions $R^{C\times H\times W}$. Subsequently, B and C are reshaped to $R^{C\times N}$, where N = H×W represents the total number of pixels.

A matrix multiplication is performed between the transpose of C and B, followed by a softmax operation to compute the spatial attention map S $\in R^{N\times N}$:

$$S_{ji} = \frac{\exp\left(B_i \cdot C_j\right)}{\sum_{i=1}^{N} \exp\left(B_i \cdot C_j\right)} \tag{3.1}$$

In this equation, S_{ji} quantifies the influence of the i-th position on the j-th position. The matrix D is then reshaped to $R^{C\times N}$ and multiplied with the transpose of S. The resulting product is reshaped back to $R^{C\times H\times W}$.

To balance the contribution of the position attention features extracted by PAM with the original features, a learnable parameter α is introduced. The final output E $\in R^{C\times H\times W}$ is obtained by multiplying the reshaped product by α and performing an element-wise sum with the original features A:

$$E_j = \alpha \sum_{i=1}^{N} (S_{ji} D_i) + A_j \tag{3.2}$$

The weight α is initialized as 0 and learned progressively during training. PAM's strong capability to extract spatial features is evident from Equation 2. The resulting feature E at each position is a weighted sum of features across all positions and the original features, indicating that it incorporates global contextual information while aggregating context based on the spatial attention map. This mechanism ensures effective extraction of position-specific features while maintaining a comprehensive global context, making PAM particularly suitable for tasks requiring fine-grained spatial understanding, such as medical image segmentation.

CAM (Channel Attention Module): As shown in Figure 3.4, this is CAM, which excels in extracting channel characteristics. Unlike PAM, this

study directly reshapes the original feature A $\in R^{C\times H\times W}$ to $R^{C\times N}$, and then performs matrix multiplication between A and its transpose. Subsequently, this study apply a softmax layer to obtain the channel attention map X $\in R^{C\times C}$:

$$X_{ji} = \frac{\exp\left(A_i \cdot A_j\right)}{\sum_{i=1}^{C} \exp\left(A_i \cdot A_j\right)} \tag{3.3}$$

Here, x_{ji} measures the impact of the i-th channel on the j-th channel. Next, this study performs matrix multiplication between the transpose of X and A, reshaping the result to $R^{C\times H\times W}$. β is a learnable parameter that controls the fusion ratio between the channel attention features and the original features. This study then multiply the result by a scale parameter β and perform an element-wise sum operation with A to obtain the final output E $\in R^{C\times H\times W}$:

$$E_j = \beta \sum_{i=1}^{N} (X_{ji} A_i) + Aj \tag{3.4}$$

Like α, β is learned through training. Similarly to PAM, during the extraction of channel features in CAM, the final feature for each channel is generated as a weighted sum of all channels and original features, thus endowing CAM with powerful channel feature extraction capabilities.

DA (Dual Attention Module): Figure 3.2 illustrates the Dual Attention Block (DA-Block) architecture, which combines the Positional Attention Module (PAM) and Channel Attention Module (CAM) to enhance feature extraction. The DA-Block consists of two main components, one focused on PAM and the other on CAM.

The first component processes input features through a convolution operation, reducing the number of channels by a factor of sixteen to obtain α^1. This step simplifies the extraction of PAM features and adjusts the feature dimensions for subsequent attention computations. After PAM processing and another convolution, $\hat{\alpha^1}$ is produced, further refining the extracted features.

Similarly, the second component applies CAM-focused processing. This dual-attention approach allows the DA-Block to capture both spatial and channel-wise dependencies effectively. By integrating positional and channel information, the DA-Block achieves a comprehensive feature extraction process, particularly beneficial for medical image segmentation tasks where both spatial relationships and channel-specific information are crucial.

$$\alpha^1 = Conv\left(input\right) \tag{3.5}$$

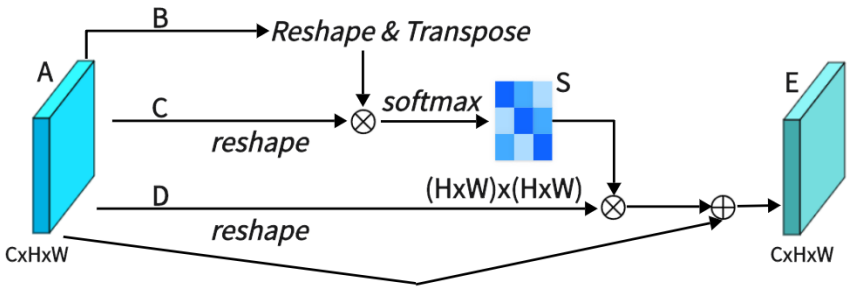

Figure 3.3: Architecture of Position Attention Mechanism(PAM).

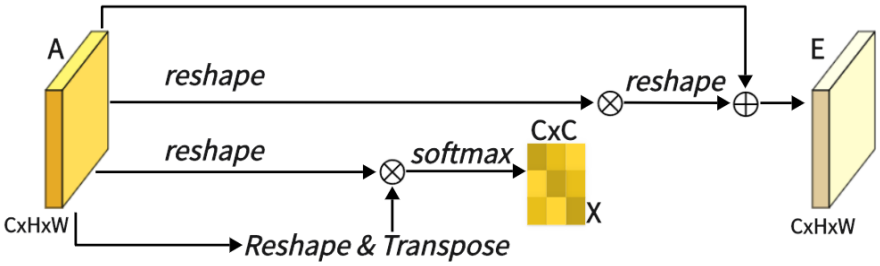

Figure 3.4: Architecture of Channel Attention Mechanism(CAM).

$$\hat{\alpha^1} = Conv\left(PAM\left(\alpha^1\right)\right) \tag{3.6}$$

The other component is the same, with the only difference being that the PAM block is replaced with a CAM with the following formula:

$$\alpha^2 = Conv\left(input\right) \tag{3.7}$$

$$\hat{\alpha^2} = Conv\left(CAM\left(\alpha^2\right)\right) \tag{3.8}$$

After extracting $\hat{\alpha^1}$ and $\hat{\alpha^2}$ from the two layers of attention, the output is obtained by aggregating and summing the two layers of attention and recovering the number of channels in one convolution.

$$output = Conv\left(\hat{\alpha^1} + \hat{\alpha^2}\right) \tag{3.9}$$

To optimize DA-Block for medical image segmentation, this study fine-tuned the number of intermediate channels. This optimization allows the model to focus on the most critical features, improving its sensitivity to key information in medical images. By adapting the DA-Block to the specific

characteristics of medical images, this study allows the model to better capture the fine-grained details necessary for accurate segmentation. This targeted optimization sets the approach apart from previous work, which often overlooks the importance of tailoring attention mechanisms to the unique demands of medical image segmentation.

3.2.3 Encoder with Transformer and Dual Attention

As illustrated in Figure 3.1, the encoder architecture consists of four key components: convolution blocks, DA-Block, embedding layers, and transformer layers. Of particular significance is the inclusion of the DA block before the transformer layer. This design is aimed at performing specialized image processing on the postconvolution features, enhancing the Transformer's feature extraction for image content. While the Transformer architecture plays a crucial role in preserving global context, the DA block strengthens the Transformer's capability to capture image-specific features, enhancing its ability to capture global contextual information in the image. This approach effectively combines global features with image-specific spatial and channel characteristics.

The first component comprises the three convolutional blocks of the architecture of the U-Net and its diverse iterations, seamlessly integrating convolutional operations with downsampling processes. Each convolutional layer halves the size of the input feature map and doubles its dimension, a configuration empirically found to maximize feature expressiveness while maintaining computational efficiency. The second component uses DA-Block extract features at both the positional and channel levels, enhancing the depth of feature representation while preserving the intrinsic characteristics of the input map. The third component is that the embedding layer serves as a critical intermediary, enabling the requisite dimensional adaptation, a prelude to the subsequent transformer strata. The fourth component integrates transformer layers for enhanced global feature extraction, beyond the reach of traditional CNNs.

Putting the above parts together, it works as follows: the input image traverses three consecutive convolutional blocks, systematically expanding the receptive field to encompass vital features. Subsequently, the DA-Block refines features through the application of both position-based and channel-based attention mechanisms. Following this, the remodeled features undergo a dimensionality transformation courtesy of the embedding stratum before they are channeled into the Transformer framework for the extraction of all-encompassing global features. This orchestrated progression safeguards the comprehensive retention of information across the continuum of successive

convolutional layers. Ultimately, the Transformer-generated feature map is restructured and navigated through skip connection layers to feed into the decoder.

By combining convolutional neural networks, transformer architectures, and dual-attention mechanisms, the encoder configuration culminates in a robust capability for feature extraction, resulting in a symbiotic powerhouse of capabilities.

3.2.4 Skip-connections with Dual Attention

Similar to other U-structured models, this study have also incorporated skip connections between the encoder and decoder to bridge the semantic gap that exists between them. To further minimize this semantic gap, this study introduced dual attention blocks (DA-blocks), as shown in Figure 3.1, in each of the three skip connection layers. This decision was based on the observation that traditional skip connections often transmit redundant features, which DA-Blocks effectively filter. Integrating DA-Blocks into the skip connections allows them to refine the sparsely encoded features from both positional and channel perspectives, extracting more valuable information while reducing redundancy. By doing so, DA-Blocks help the decoder to reconstruct more accurate feature maps. Moreover, the inclusion of DA-Blocks not only enhances the model's robustness but also effectively mitigates sensitivity to overfitting, contributing to the overall performance and generalization capability of the model.

3.2.5 Decoder

As depicted in Figure 3.1, the right half of the diagram corresponds to the decoder. The primary role of the decoder is to reconstruct the original feature map utilizing features acquired from the encoder and those received through skip connections, employing operations like upsampling.

The decoder's components include feature fusion, a segmentation head, and three upsampling convolution blocks. The first component: feature fusion involves the integration of feature maps transmitted through skip connections with the existing feature maps, thereby assisting the decoder in faithfully reconstructing the original feature map. The second component: the segmentation head is responsible for restoring the final output feature map to its original dimensions. The third component: the three upsampling convolution blocks incrementally double the size of the input feature map in each step, effectively restoring the image's resolution.

Putting the above parts together, the workflow begins by passing the input image through convolution blocks and subsequently performing upsampling to augment the size of the feature maps. These feature maps undergo a two-fold size increase, while their dimensions are reduced by half. The features received through the skip connections are then fused, followed by continued upsampling and convolution. After three iterations of this process, the generated feature map undergoes a final round of upsampling and is accurately restored to its original size by the segmentation head.

Thanks to this architecture, the decoder demonstrates robust decoding capabilities, effectively revitalizing the original feature map using features from both the encoder and skip connections.

Furthermore, compared to other transformer-based approaches that extensively utilize transformer blocks throughout the architecture, such as Swin-Unet, DA-TransUNet achieves a more favorable balance between performance and computational efficiency. The judicious integration of DA-Blocks in the encoder and skip connections allows DA-TransUNet to enhance feature representation while maintaining a manageable computational footprint.

3.3 Experimental

To assess the efficacy of the proposed DA-TransUNet model, comprehensive experiments were carried out in six diverse medical imaging datasets: Synapse [57], CVC-ClinicDB [58], Chest X-ray Masks and Labels [63, 64], Kvasir-SEG [61], Kvasir-Instrument [62], and ISIC 2018 Task [59, 60]. The results of these experiments demonstrate the superior performance of DA-TransUNet compared to existing state-of-the-art methods across all evaluated datasets.

The following subsections provide a detailed overview of each dataset, followed by a description of the implementation specifics and a comprehensive analysis of the results obtained for each of the six datasets.

3.3.1 Dataset Descriptions

3.3.1.1 Synapse Multi-organ Segmentation Dataset

The Synapse dataset is a comprehensive collection of abdominal CT scans, covering 30 volumetric images that capture eight distinct abdominal organs. These organs include the bilateral kidneys (left and right), the aorta, spleen, gallbladder, liver, stomach, and pancreas. In total, the dataset comprises

3,779 axially enhanced abdominal CT image slices, providing a rich resource for multiorgan segmentation tasks in the abdominal region.

3.3.1.2 CVC—ClinicDB

CVC-ClinicDB is a database of frames extracted from colonoscopy videos, which is part of the Endoscopic Vision Challenge. This is a dataset of endoscopic colonoscopy frames for the detection of polyps. CVC-ClinicDB contains 612 still images from 29 different sequences. Each image has its associated manually annotated ground truth covering the polyp.

3.3.1.3 Chest Xray

Chest Xray Masks and Labels X-ray images and the corresponding masks are provided. X-rays were obtained from the Montgomery County Department of Health and Human Services Tuberculosis Control Program, Montgomery County, Maryland, USA. The set of images contains 80 anterior and posterior X-rays, of which 58 X-rays are normal and 1702 X-rays are abnormal with evidence of tuberculosis. All images have been de-identified and presented in DICOM format. The set contains a variety of abnormalities, including exudates and corneal morphology. It contains 138 posterior-anterior radiographs, of which 80 radiographs were normal and 58 radiographs showed abnormal manifestations of tuberculosis.

3.3.1.4 Kvasir SEG

Kvasir SEG is an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated and verified by an experienced gastroenterologist. It contains 1000 polyp images and their corresponding ground truths, the resolution of the images contained in Kvasir-SEG varies from 332x487 to 1920x1072 pixels, and the file format is jpg.

3.3.1.5 Kvasir-Instrument Dataset

The Kvasir-Instrument dataset is a specialized collection focused on gastrointestinal endoscopic instruments. This comprehensive dataset comprises 590 high-quality endoscopic images, each accompanied by its corresponding ground truth segmentation mask. The images in this collection showcase a diverse array of gastrointestinal (GI) procedure instruments, including snares, balloons, biopsy forceps, and other essential tools used in endoscopic procedures.

Image resolutions within the dataset vary, ranging from 720x576 to 1280x1024 pixels, providing a realistic representation of the variability encountered in clinical settings. All images are stored in the widely-used JPEG format, ensuring compatibility with most image processing software. This carefully curated collection serves as a valuable resource for developing and evaluating algorithms aimed at instrument detection and segmentation in endoscopic procedures. Such applications are crucial for advancing computer-assisted interventions in gastroenterology, potentially improving the accuracy and efficiency of diagnostic and therapeutic endoscopic procedures.

3.3.1.6 2018ISIC-Task

The dataset used in the 2018 ISIC Challenge addresses the challenges of skin diseases. It comprises a total of 2512 images, with a file format of JPG. The images of lesions were obtained using various dermatoscopic techniques from different anatomical sites (excluding mucous membranes and nails). These images are sourced from historical samples of patients undergoing skin cancer screening at multiple institutions. Each lesion image contains only a primary lesion.

3.3.2 Implementation Settings

3.3.2.1 Baselines

To establish the efficacy of our proposed approach in the domain of medical image segmentation, we conducted a comprehensive evaluation against a diverse set of established and state-of-the-art models. The selection of comparative models encompasses both foundational architectures and recent innovations in the field.

Our evaluation begins with the seminal U-Net [51], which has served as a cornerstone in biomedical image segmentation. We then consider its advanced variants: UNet++ [18], which enhances the original architecture with dense skip connections and deep supervision; and Attention U-Net [6], which incorporates attention gates for more precise feature selection.

We further extend our comparison to include more recent developments: DA-UNet [30], which leverages dual attention mechanisms to enrich feature extraction, and TransUNet [9], which integrates transformer modules to capture global context effectively.

To ensure a thorough assessment against the latest advancements, we also include several cutting-edge models in our benchmark:

- UCTransNet [65], which innovatively applies attention mechanisms to skip connections within the U-Net framework
- TransNorm [66], which incorporates transformer modules in both the encoder and skip connections of the standard U-Net
- MIM [67], featuring a novel transformer-based design specifically tailored for medical image segmentation tasks

Through this extensive comparison against both well-established baselines and advanced contemporary models, our aim is twofold: to demonstrate the unique strengths of our proposed approach and to highlight its potential for wide-ranging applications in medical image analysis. This comprehensive evaluation serves to position our model within the current landscape of medical image segmentation techniques, showcasing its superior performance and innovative features.

3.3.2.2 Implementation Details

The proposed DA-TransUNet model was implemented using the PyTorch deep learning framework and trained on an NVIDIA RTX 3090 GPU [68]. The training process incorporated the following key parameters and considerations:

For most data sets, the model was configured with an input image resolution of 256x256 pixels and a patch size of 16. The optimization process used the Adam algorithm with the following hyperparameters: a learning rate of 1×10^{-3}, a momentum of 0.9, and a weight decay coefficient of 1×10^{-4}.

The training duration was set to 500 epochs for most datasets. However, to account for the varying sizes of different datasets and ensure convergence, the training process for the chest X-ray masks and labels dataset, as well as the ISIC 2018-Task dataset, was adjusted to 50 epochs.

During the training phase on five datasets, including CVC-ClinicDB, the DA-TransUNet model was trained end to end. The objective function was formulated as a combination of weighted binary cross-entropy (BCE) and Dice coefficient loss:

$$\text{Loss}_{\text{combined}} = 0.5 \times \text{BCE} + 0.5 \times \text{DiceLoss} \tag{3.10}$$

For the Synapse dataset, to ensure a fair evaluation, we employed the pre-trained "R50-ViT" model with an adjusted input resolution of 224x224 pixels and a patch size of 16. The optimization process for this dataset utilized the SGD algorithm with a learning rate of 0.01, momentum of 0.9, and weight decay of 1×10^{-4}. The batch size was set to 24. The loss function for the Synapse dataset was defined as:

$$\text{Loss}_{\text{Synapse}} = 0.5 \times \text{CrossEntropyLoss} + 0.5 \times \text{DiceLoss} \tag{3.11}$$

These carefully tuned parameters and dataset-specific adjustments were crucial in optimizing DA-TransUNet performance across various medical image segmentation tasks.

This loss function balances the contributions of cross-entropy and Dice losses, ensuring impartial evaluation during testing on the synapse dataset.

When using the data sets, this study uses a 3 to 1 ratio, where 75% is the training set and 25% is the test set, to ensure the adequacy of training.

3.3.2.3 Model Evaluation

In evaluating the performance of DA-TransUNet, this study uses a comprehensive set of metrics including Intersection over Union (IoU), Dice Coefficient(DSC), and Hausdorff Distance (HD). These metrics are industry standards in computer vision and medical image segmentation, providing a multifaceted assessment of the model's accuracy, precision, and robustness.

The choice of these metrics is based on their complementary nature and the ability to capture different aspects of segmentation quality. IoU and DSC measure the overlap between the predicted and ground truth segmentation masks, providing a global assessment of the model's ability to accurately identify and delineate target structures. HD, on the other hand, captures the maximum distance between the predicted and ground truth segmentation boundaries, ensuring that the predicted segmentation closely adheres to the true boundaries of the target structures, even in the presence of small segmentation errors or irregularities.

IOU (Intersection over Union) is one of the commonly used metrics to evaluate the performance of computer vision tasks such as object detection, image segmentation, and instance segmentation. Measures the degree of overlap between the predicted region of the model and the actual target region, helping us to understand the accuracy and precision of the model. In target detection tasks, IOU is usually used to determine the degree of overlap between the predicted bounding box (Bounding Box) and the real bounding box. In image segmentation and instance segmentation tasks, IOU is used to evaluate the degree of overlap between the predicted region and the ground-truth segmentation region.

$$IOU = \frac{TP}{FP + TP + FN} \tag{3.12}$$

The Dice coefficient (also known as the Srensen-Dice coefficient, F1 score, DSC) is a measure of model performance in image segmentation tasks and is particularly useful for dealing with class imbalance problems. Measures the degree of overlap between the predicted results and the ground-truth segmentation results, and is particularly effective when dealing with segmentation of objects with unclear boundaries. The Dice coefficient is commonly used as a measure of the model's accuracy on the target region in image segmentation tasks and is particularly suitable for dealing with relatively small or uneven target regions.

$$\text{Dice}(P, T) = \frac{|P_1 \cap T_1|}{|P_1| + |T_1|} \Leftrightarrow \text{Dice} = \frac{2|T \cap P|}{|F| + |P|} \tag{3.13}$$

The Hausdorff Distance (HD) serves as a crucial metric in the evaluation of image segmentation models, particularly in the domain of medical imaging. This measure quantifies the degree of similarity between two point sets, making it especially valuable for assessing the accuracy of segmentation boundaries. In the context of medical image analysis, HD provides a robust means of comparing predicted segmentation outputs against ground truth annotations.

The fundamental principle of the Hausdorff Distance lies in its ability to capture the maximum discrepancy between two segmentation contours. Mathematically, it can be expressed as:

$$H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a)\} \tag{3.14}$$

where A and B represent the two sets of points being compared (e.g. predicted and true segmentation boundaries), and $d(a, b)$ denotes the distance between points a and b.

The distinctive feature of HD is its sensitivity to outliers, as it identifies the most significant disparity between the two sets. This characteristic makes HD particularly adept at evaluating segmentation performance in regions where precise boundary delineation is critical, such as in tumor margin detection or organ boundary identification in medical imaging.

By providing a quantitative measure of the maximum deviation between predicted and true segmentations, HD offers valuable insights into a model's ability to accurately capture intricate boundary details. This property is especially beneficial in medical applications where even small inaccuracies in segmentation boundaries can have significant clinical implications.

This study evaluate using both Dice and HD in the Synapse dataset and both Dice and IOU in other datasets.

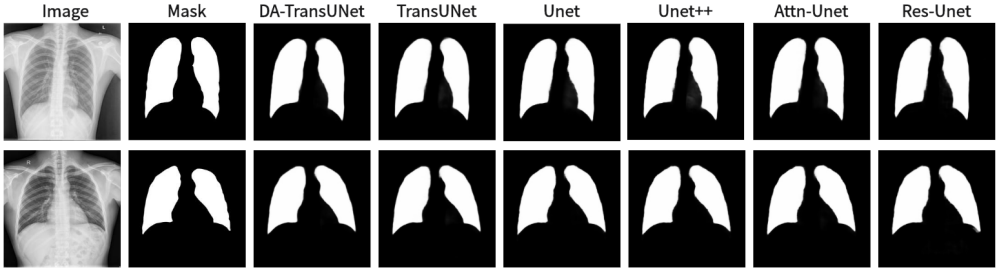

Figure 3.5: Comparison of qualitative results between DA-TransUNet and existing models on the task of segmenting Chest X-ray Masks and Labels X-ray datasets.

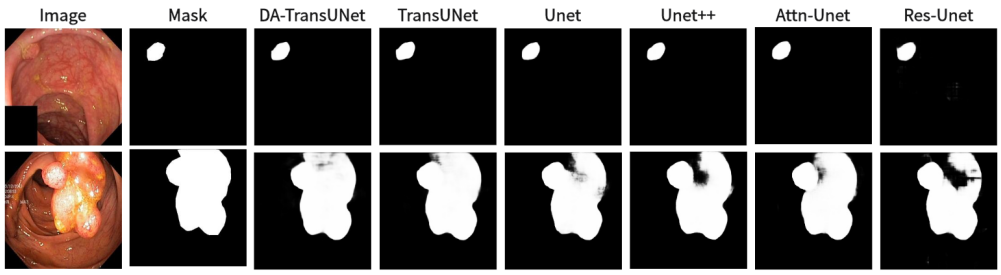

Figure 3.6: Comparison of qualitative results between DA-TransUNet and existing models on the task of segmenting Kvasir-Seg datasets.

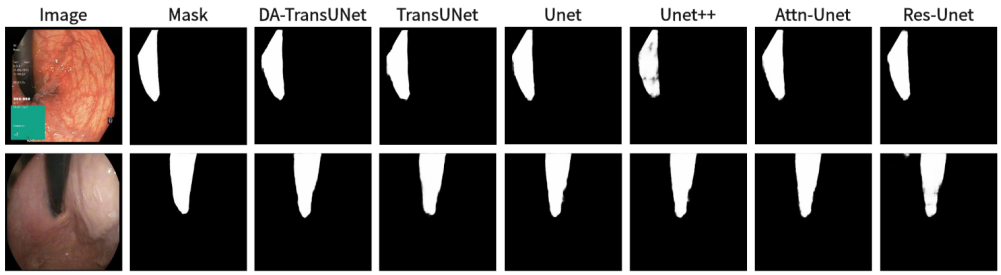

Figure 3.7: Comparison of qualitative results between DA-TransUNet and existing models on the task of segmenting Kavsir-Instrument datasets.

3.3.3 Comparison to the State-of-the-Art Methods

3.3.3.1 Segmentation Performance and Comparison

To rigorously evaluate the performance of our proposed DA-TransUNet model, we conducted a comprehensive comparative analysis against a spec-

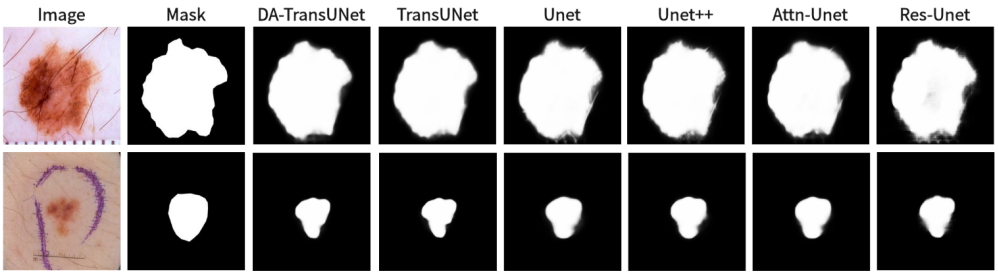

Figure 3.8: Comparison of qualitative results between DA-TransUNet and existing models on the task of segmenting 2018ISIC-Task datasets.

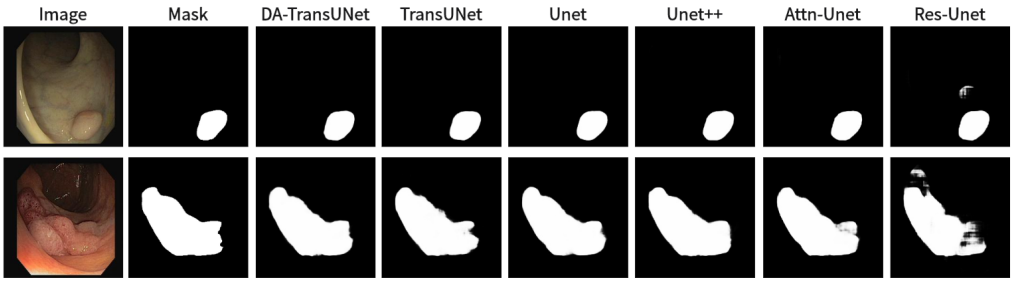

Figure 3.9: Comparison of qualitative results between DA-TransUNet and existing models on the task of segmenting CVC-ClinicDB datasets.

trum of state-of-the-art (SOTA) segmentation models. The benchmark models selected for this comparison include the foundational U-Net [51], as well as its advanced variants such as Res-UNet [5], U-Net++ [18], and Att-UNet [6]. Furthermore, we included more recent architectures that leverage transformer mechanisms, namely TransUNet [9], TransNorm [66], UCTransNet [65], and Swin-UNet [11]. The comparison also encompasses other innovative designs like MultiResUNet [69] and MIM [67].

The primary experimental evaluation was performed on the Synapse multiorgan segmentation dataset, with the results presented in Table 3.1. This dataset was chosen for its complexity and relevance in medical image analysis.

As illustrated in Figure 3.11, our DA-TransUNet model achieved remarkable performance metrics, with an average Dice Similarity Coefficient (DSC) of 79.80% and an average Hausdorff Distance (HD) of 23.48 mm. These results represent significant improvements of 2.32% in DSC and 8.21 mm in

HD compared to the TransUNet baseline. Such enhancements indicate the superior capability of DA-TransUNet in both overall segmentation accuracy and precise organ boundary delineation.

Figure 3.10 further corroborates these findings, showing that DA-TransUNet achieves the highest DSC value among all the compared models. Although its HD performance is marginally higher than Swin-UNet, it still shows substantial improvement over several recent models, including TransUNet.

In terms of computational efficiency, DA-TransUNet requires 35.98 ms for segmenting a single image, compared to 33.58 ms for TransUNet. This minimal difference in processing time suggests that the improved segmentation quality of DA-TransUNet comes with a negligible additional computational cost.

Detailed analysis of individual organ segmentation reveals that DA-TransUNet outperforms TransUNet across multiple organs:

- Gallbladder: 2.14% improvement
- Right kidney: 3.43% improvement
- Liver: 0.48% improvement
- Spleen: 3.45% improvement
- Stomach: 4.11% improvement
- Pancreas: A notable 5.73% improvement

While DA-TransUNet demonstrates superior performance in most organs, it shows slight decreases in segmentation accuracy for the aorta (0.69%) and left kidney (0.17%) compared to TransUNet. Nevertheless, the model achieves peak segmentation rates for the right kidney, liver, pancreas, and stomach, indicating its enhanced feature learning capabilities for these specific anatomical structures.

These comprehensive results underscore the efficacy of DA-TransUNet in medical image segmentation tasks, particularly in scenarios requiring high precision and robust performance across diverse anatomical structures.

To further confirm the better segmentation of our model compared to TransUNet, this study visualized the segmentation plots of TransUNet and DA-TransUNet (see Figure3.11). From the yellow and purple parts in the first column, this study can see that our segmentation effect is obviously better than that of TransUNet; from the second column, the extension of purple is better than that of TransUNet, and there is no vacancy in the blue part; from the third column, there is a semicircle in the yellow part, and the vacancy in red is smaller than that of TransUNet, etc. It is evident that DA-TransUNet outperforms TransUNet in segmentation quality. In summary, DA-TransUNet significantly surpasses TransUNet in segmenting

the left kidney, right kidney, spleen, stomach, and pancreas. It also offers superior visualization performance in image segmentation.

This study simultaneously took DA-TransUNet in five datasets, CVC-ClinicDB, chest X-ray masks and labels, ISIC2018-Task, kvasir instrument and kvasir-seg, and compared it with some classical models (see Table 3.2). In the table, the values of IOU and Dice of DA-TransUNet are higher than those of TransUNet in the five datasets, CVC-ClinicDB, Chest X-ray Masks and Labels, ISIC2018-Task, kvasir-instrument and kvasir-seg. In addition, DA-TransUNet has the best dataset segmentation in four of the five datasets. As seen in the table, our DA-TransUNet has more excellent feature learning and image segmentation capabilities.

This study also shows the results of the image segmentation visualization of DA-TransUNet in these five datasets, and this study also show the results of the comparison models for the comparison. The visualization results for chest X-ray masks and labels, Kvasir-Seg, Kvasir-Instrument, ISIC2018-Task and CVC-ClinicDB datasets are presented in Figure3.5, Figure3.6, Figure3.7, Figure3.8, and Figure3.9, respectively. In the figure, it can be seen that the segmentation effect of DA-TransUNet has a good performance. Firstly, DA-TransUNet has better segmentation results than TransUNet. In addition, compared to the four classical models of U-net, Unet++, Attn-Unet, and Res-Unet, DA-TransUNet has a certain improvement. It can be seen that the effectiveness of DA-TransUNet for model segmentation is confirmed not only in the Synapse dataset, but also in the five datasets (CVC-ClinicDB, Chest X-ray Masks and Labels, ISIC2018-Task, kvasir-instrument, kvasir-seg). This study further establishes that DA-TransUNet excels in both 3D and 2D medical image segmentation.

Table 3.1: Experimental results on the Synapse dataset

Model	Year	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Sple en	Stomach
U-net [51]	2015	76.85%	39.70	89.07	**69.72**	77.77	68.6	93.43	53.98	86.67	75.58
U-Net++ [18]	2018	76.91%	36.93	88.19	68.89	81.76	75.27	93.01	58.20	83.44	70.52
Residual U-Net [5]	2018	76.95%	38.44	87.06	66.05	**83.43**	76.83	93.99	51.86	85.25	70.13
Att-Unet [6]	2018	77.77%	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
MultiResUNet [69]	2020	77.42%	36.84	87.73	65.67	82.08	70.43	93.49	60.09	85.23	75.66
TransUNet [9]	2021	77.48%	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet [65]	2022	78.23%	26.75	84.25	64.65	82.35	77.65	94.36	58.18	84.74	79.66
TransNorm [66]	2022	78.40%	30.25	86.23	65.1	82.18	78.63	94.22	55.34	89.50	76.01
MIM [67]	2022	78.59%	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
swin-unet [11]	2022	79.13%	**21.55**	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
DA-TransUNet(Ours)	2023	**79.80%**	23.48	86.54	65.27	81.70	**80.45**	**94.57**	**61.62**	88.53	**79.73**
Average Relative Improvement	-	**2.03**%	**-9.00**	**-0.73**%	**-1.09**%	**0.28**%	**5.21**%	**0.82**%	**4.86**%	**1.97**%	**4.5**%

3.3.3.2 Computational Complexity and Efficiency

The integration of DA-Blocks in the encoder and skip connections introduces additional computational overhead compared to the standard TransUNet

Table 3.2: Experimental results of datasets (CVC-ClinicDB, Chest Xray Masks and Labels, ISIC2018-Task, kvasir-instrument, kvasir-seg)

	CVC-ClinicDB		Chest Xray Masks and Labels		ISIC2018-Task		kvasir-instrument		kvasir-seg	
	Iou↑	Dice↑	Iou↑	Dice↑	Iou↑	Dice↑	Iou↑	Dice↑	Iou↑	Dice↑
U-net [51]	0.7821	0.8693	0.9303	0.9511	0.8114	0.8722	0.8957	0.9358	0.8012	0.8822
Attn-Unet [6]	0.7935	0.8741	0.9274	0.9503	0.8151	0.876	0.8949	0.9359	0.7801	0.8661
Unet++ [18]	0.7847	0.8714	0.9289	0.9505	0.8133	0.873	**0.8995**	**0.9389**	0.7767	0.8657
ResUNet [5]	0.5902	0.7422	0.9262	0.9505	0.7651	0.8332	0.8572	0.9141	0.6604	0.7785
TransUNet [9]	0.8163	0.8901	0.9301	0.9535	0.8263	0.8878	0.8926	0.9363	0.8003	0.8791
DA-TransUNet(Ours)	**0.8251**	**0.8947**	**0.9317**	**0.9538**	**0.8278**	**0.8888**	0.8973	0.9381	**0.8102**	**0.8847**

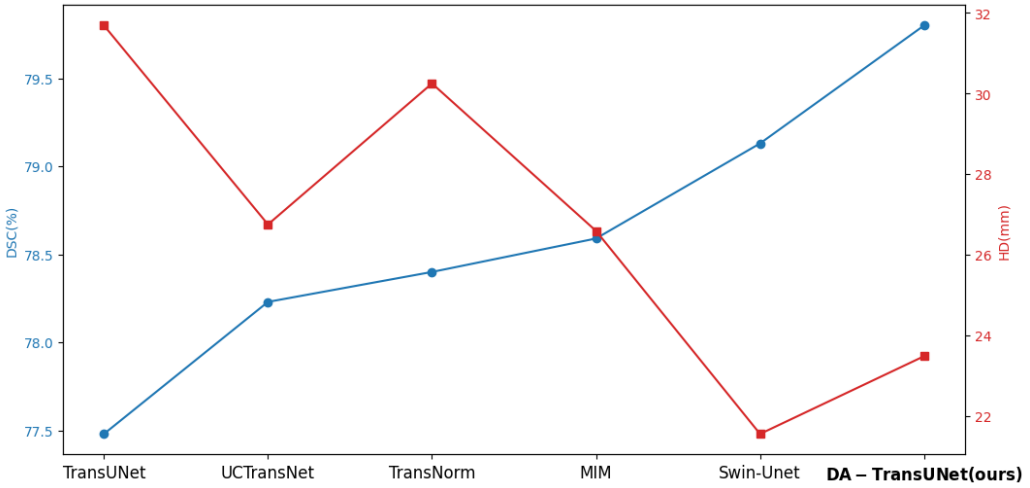

Figure 3.10: Line chart of DSC and HD values of several advanced models in the Synapse dataset

architecture. Let the input feature map have a spatial resolution of $H \times W$ and C channels. The computational complexity of the Position Attention Module (PAM) is $\mathcal{O}(H^2W^2C)$, while the Channel Attention Module (CAM) has a complexity of $\mathcal{O}(C^2HW)$. As the DA-Block consists of both PAM and CAM, its overall computational complexity is $\mathcal{O}(H^2W^2C + C^2HW)$. However, it is worth noting that the DA-Block itself is not computationally intensive, as it only involves simple matrix multiplications and element-wise operations.

Table 3.3 compares the number of parameters, Dice Similarity Coefficient (DSC), and Hausdorff Distance (HD) between DA-TransUNet and TransUNet. Incorporation of DA-Blocks leads to a modest increase of 2.54% in the number of parameters compared to TransUNet. This incremental increase in parameters is justifiable considering the substantial performance gains achieved by DA-TransUNet, as demonstrated in our experimental results.

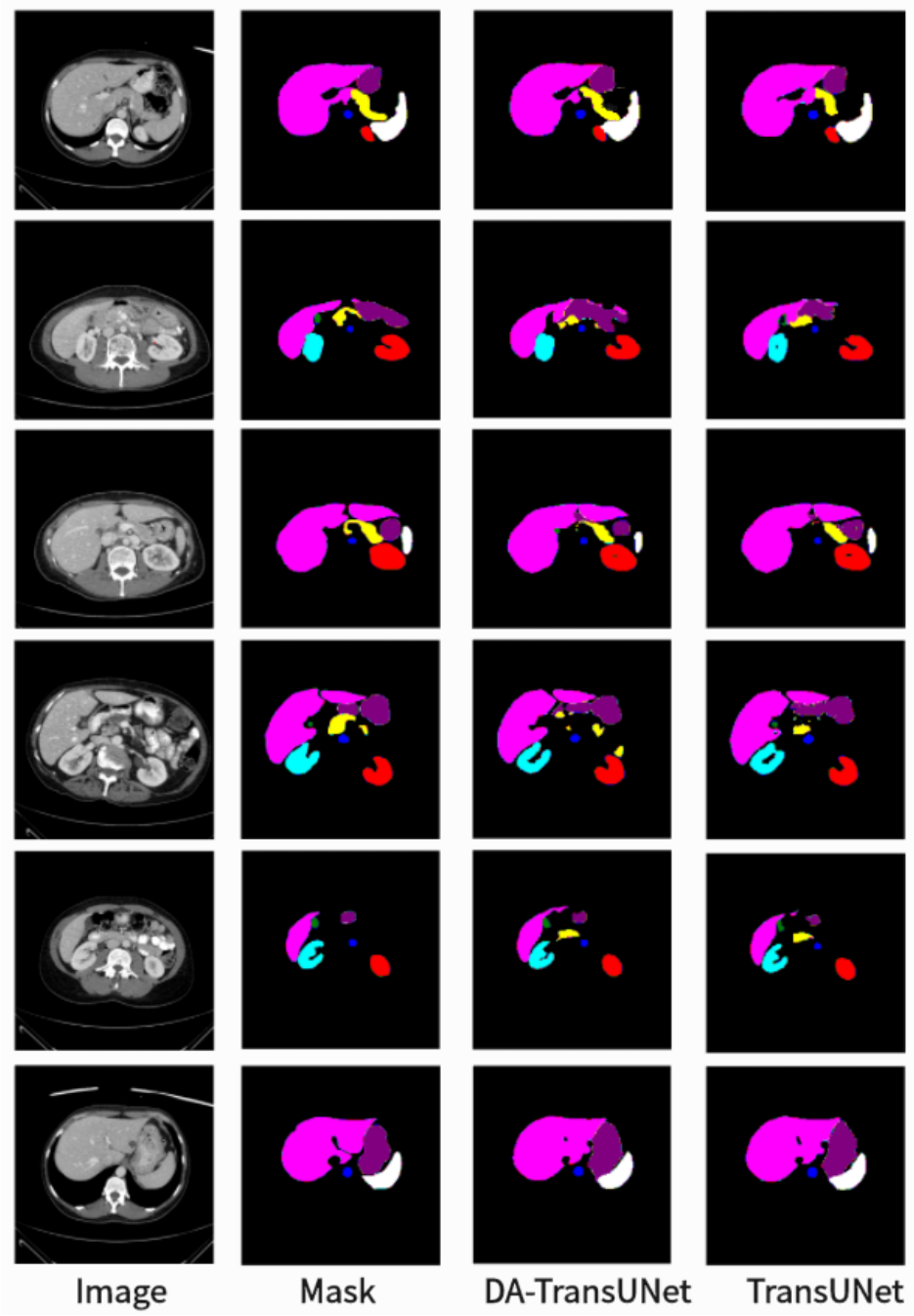

Figure 3.11: Segmentation results of TransUNet and DA-TransUNet on the Synapse dataset.

DA-TransUNet achieves an average improvement of 2.99% in DSC and 25.9%

in HD compared to TransUNet. The strategic placement of DA-Blocks allows for efficient feature refinement while maintaining a reasonable model size.

Table 3.3: Comparison of model parameters and performance between DA-TransUNet and TransUNet.

Model	Params	Params Increase	DSC Improvement	HD Improvement
TransUNet	105,276,066	-	-	-
DA-TransUNet	**107,950,840**	**2.54%**	**2.99%**	**25.9%**

3.3.4 Ablation Study

This study conducted ablation experiments on the DA-TransUNet model using the Synapse data set to discuss the effects of different factors on model performance. Specifically, it includes: 1) DA-Block in Encoder. 2) DA-Block in Skip Connection.

3.3.4.1 Effect of the DA-Block in Encoder And Skip Connection

This study conducted a series of experiments to evaluate the effectiveness of integrating DA-Blocks into various components of the model architecture, as illustrated in Table 3.4. The investigation focused on two key areas: the incorporation of DA-Blocks into skip connections and their placement within the encoder.

The introduction of DA-Blocks at each layer of skip connections yielded notable improvements in model performance. Specifically, the Dice Similarity Coefficient (DSC) increased from a baseline of 77.48% to 78.28%, while the Hausdorff Distance (HD) metric showed a reduction from 31.69mm to 29.09mm. These results suggest that the DA-Blocks enhance feature refinement in skip connections, potentially mitigating information loss during upsampling and contributing to improved model stability and reduced overfitting.

Further experimentation involved placing DA-Blocks in the encoder, preceding the Transformer layer. This modification resulted in an even more substantial improvement, with the DSC rising to 78.87% and the HD decreasing to 27.71mm. The marked enhancement in both metrics underscores the significance of feature refinement prior to transformer processing.

The cumulative findings, as presented in Table 3.4, provide strong evidence for the efficacy of DA-Blocks in medical image segmentation. The strategic placement of these blocks, both within skip connections and before the transformer layer in the encoder, demonstrates a synergistic effect

that significantly enhances the model's overall segmentation capabilities for medical imaging tasks.

Table 3.4: Effects of Combinatorial Placement of DA-Blocks in the Encoder and Through Skip Connections on Performance Metrics

	Encoder with DA	Skip with DA	DSC↑	HD↓
DA-TransUNet			77.48	31.69
DA-TransUNet		√	78.28	29.09
DA-TransUNet	√		78.87	27.71
DA-TransUNet	√	√	**79.80**	**23.48**

Table 3.5: Effects of Incorporating DA-Block in the Encoder and Skip Connections at Different Layers on Performance Metrics

	1st layer	2nd layer	3rd layer	DSC↑	HD↓
DA-TransUNet				78.87	27.71
DA-TransUNet	√			79.36	25.80
DA-TransUNet		√		78.65	23.43
DA-TransUNet			√	79.49	30.71
DA-TransUNet	√	√	√	**79.80**	**23.48**

Table 3.6: Effect of the number of intermediate channels in DA-Block

	1	2	4	8	16	32	DSC↑	HD↓
DA-TransUNet	√						78.55	28.22
DA-TransUNet		√					79.35	23.77
DA-TransUNet			√				79.71	25.90
DA-TransUNet				√			79.35	25.66
DA-TransUNet					√		**79.80**	**23.48**
DA-TransUNet						√	79.71	24.45

3.3.4.2 Effect of adding DA-Blocks to skip connections in different layers

Based on the quantitative results of Table 3.5, this study experimented with various configurations of the placement of the DA block in three different layers of skip connections to identify the optimal architectural layout to enhance the performance of the model. Specifically, when DA blocks were added to just the first layer, the DSC metric improved to 79.36% from a

baseline of 78.87%, and the HD metric decreased to 25.80mm from 27.71mm. Adding DA-Blocks to the second and third layers resulted in some progress. When DA locks were integrated across all layers, there was an improvement, reflected by a DSC of 79.80% and a HD of 23.48mm. In contrast to traditional architectures where skip connections indiscriminately pass features from the encoder to the decoder, our approach with DA-Blocks selectively improves feature quality at each layer. The results, as corroborated by Table 3.5, reveal that introducing DA blocks into even a single layer improves performance, and the greatest gains are observed when applied across all layers. This indicates the effectiveness of integrating DA-Blocks within skip connections to enhance both feature extraction and medical image segmentation. Therefore, the table clearly supports the idea that the layer-wise inclusion of DA-Blocks in skip connections is an effective strategy to improve medical image segmentation.

3.3.4.3 Effect of the number of intermediate channels in DA-Block

Based on the results shown in Table3.6, this study conducted a discussion on the size of the intermediate layer in the DA-Block, which demonstrates the effectiveness of convolutional layers from an experimental perspective. The original DA-Block had an intermediate layer size that is one-fourth of the input layer size. However, since its intended application is for road scene segmentation and not specifically tailored for medical image segmentation, this study deemed that setting the intermediate layer size at one-fourth of the input layer size might not be suitable for the medical image segmentation domain. As seen in the graph, when this study set the intermediate layer size to be the same as the input size, the evaluation results show a DSC of 78. 55% and a HD of 28.22 mm. In the related DANet research [54], where the intermediate layer was set to one-fourth of the input layer, the DSC result was 79.71%, and HD was 25.90 mm. However, when this study further reduced the size of the intermediate layer to one-sixteenth of the input layer size, this study observed an improvement in DSC to 79.80%, and HD decreased to 23.48 mm. It is evident that setting the intermediate layer to one-sixteenth of the input layer size is more suitable for medical image segmentation tasks. The reduction in the intermediate layer size can help the model mitigate the risk of overfitting, optimize computational resources, and, given the precision requirements of medical image segmentation tasks, enable the model to focus more on selecting the most crucial features, thereby enhancing sensitivity to critical information for the task.

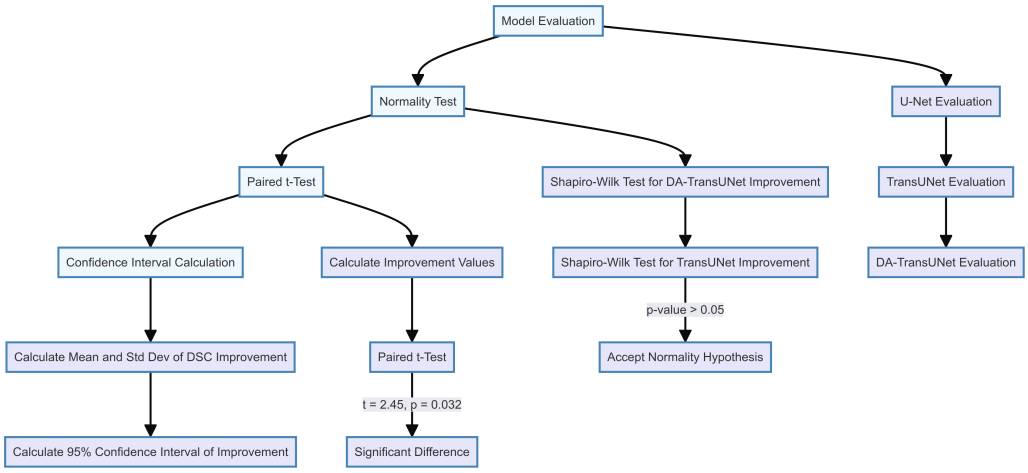

Figure 3.12: The flowchart of statistical detection is shown in the figure.

3.4 Discussion

In this present study, this study has discovered promising outcomes from the integration of DA-Blocks with the transformer and their combination with skip-connections. Encouraging results were consistently achieved across all six experimental datasets.

3.4.1 Statistical Validation of the Improvements by DA-TransUNet

To enhance the credibility of our results and further validate the superiority of DA-TransUNet, this study evaluated the performance of the models discussed in the Experimental Section (U-Net, TransUNet and DA-TransUNet) on 12 subsets of the Synapse dataset, constituting 40% of the total data, and obtained their Dice Similarity Coefficients (DSC). It is important to note that both DA-TransUNet and TransUNet are based on the U-Net architecture, which serves as a baseline model. Therefore, using U-Net as the benchmark to assess whether the improvements of DA-TransUNet over TransUNet are significant is a valid approach.

As shown in Figure.3.12, this study first assessed the normality of DSC improvement values for both DA-TransUNet and TransUNet relative to U-Net using the Shapiro-Wilk test. The results showed p-values of 0.36 and 0.82 for the improvements of DA-TransUNet and TransUNet, respectively. Since both p-values are greater than 0.05, this study cannot reject the null

hypothesis of normality. This indicates that the DSC improvement values for both DA-TransUNet and TransUNet relative to U-Net can be considered approximately normally distributed. This study then performed a paired t-test to compare the significance of the improvements. As shown in Table 3.7, the test yielded a t-statistic of 2.45 and a p-value of 0.032, demonstrating a significant difference between the improvements achieved by DA-TransUNet and TransUNet.

Furthermore, to further quantify the superiority of DA-TransUNet over TransUNet, this study calculated the confidence interval 95% for the difference in improvements between DA-TransUNet and TransUNet. The results showed that the mean difference was 3.96, with a standard deviation of 5.61, and the confidence interval was [0.40, 7.53]. This means that, at a confidence level 95%, the magnitude of the difference in DSC improvements between DA-TransUNet and TransUNet lies between 0.40 and 7.53.

To provide a comprehensive overview of the performance of the models, this study calculated the confidence intervals 95% for their DSC scores. DA-TransUNet achieved a mean DSC of 79.80 ± 5.01, with a confidence interval of [74.79, 84.81], while TransUNet achieved a mean DSC of 75.84 ± 6.77, with a confidence interval of [69.06, 82.61]. These results, summarized in Table 3.7, suggest that DA-TransUNet not only achieves higher average performance but also exhibits more consistent results compared to TransUNet.

Statistical analysis, confidence intervals, and quantification of relative improvement provide strong evidence for the superiority of DA-TransUNet over TransUNet in the task of medical image segmentation. These results highlight the effectiveness of our proposed approach and its potential to advance the field of medical image analysis.

Table 3.7: Statistical Analysis of DSC Improvements and Model Performance

Model	**Mean DSC ± SD**	**95% CI for DSC**
DA-TransUNet	79.80 ± 5.01	[74.79, 84.81]
TransUNet	75.84 ± 6.77	[69.06, 82.61]

Comparison of DSC Improvements Achieved by DA-TransUNet and TransUNet Relative to U-Net

Metric	**Mean Difference**	**95% CI for Difference**	**t-Test p-value**
Improvement	3.96	[0.40, 7.53]	0.032

3.4.2 Enhancing Feature Extraction and Segmentation with DA-Blocks

The empirical findings presented in Table 3.4 demonstrate the significant impact of integrating DA-Blocks into the encoder, notably enhancing feature

extraction capabilities and segmentation performance.

In the field of computer vision, Vision Transformers (ViT) have gained recognition for their robust global feature extraction [53]. However, they exhibit limitations in specialized tasks like medical image segmentation, where image-specific feature attention is crucial. To address this, DA-TransUNet strategically positions DA-Blocks before the Transformer module. These DA-Blocks are designed to initially extract and refine image-specific features, including spatial and channel attributes. This refined data is then processed by the Transformer, leading to enhanced global feature extraction and significantly improved feature learning and segmentation performance.

The strategic placement of DA-Blocks preceding the transformer layer represents an innovative approach that substantially improves both feature extraction efficacy and medical image segmentation precision.

Furthermore, Table 3.5 illustrates that integrating DA-Blocks with skip connections markedly enhances semantic continuity and the decoder's ability to reconstruct accurate feature maps. While traditional U-Net architectures [51] use skip connections to bridge the encoder-decoder semantic gap, our novel incorporation of Dual Attention Blocks within these layers yields promising outcomes. This integration across skip-connection layers enables focus on relevant features while filtering out extraneous information, resulting in a more efficient and accurate image reconstruction process.

The inclusion of DA-Blocks in skip connections thus represents a groundbreaking approach that enhances both feature extraction and overall model performance in medical image segmentation.

Our comprehensive evaluation across six diverse medical image segmentation datasets underscores the effectiveness and generalizability of DA-TransUNet. Consistent improvements over state-of-the-art methods, as evidenced in Table 3.1, highlight the impact of our targeted DA-Block integration. Additionally, ablation studies (Section 3.3.4) provide valuable insights into the individual contributions of DA-Blocks in various architectural components.

These findings not only emphasize the novelty of our approach but also illuminate the importance of strategically incorporating attention mechanisms for enhanced medical image segmentation. DA-TransUNet marks a significant advancement in leveraging attention mechanisms and transformers for accurate and robust segmentation across diverse medical imaging modalities. Our work opens avenues for further exploration of targeted attention mechanisms in medical image analysis, with potential implications for clinical decision-making and patient care.

3.4.3 Limitations and Future Directions

Despite the advantages, our model also has some limitations. Firstly, the introduction of the DA-Blocks contributes to an increase in computational complexity. This added cost could potentially be a hindrance in real-time or resource-constrained applications. Although this increase in parameters is relatively modest considering the performance gains achieved, it could still be a concern in resource-constrained scenarios or when dealing with very large-scale datasets. Secondly, the decoder part of our model retains the original U-Net architecture. Although this design choice preserves some of the advantages of U-Net, it also means that the decoder has not been specifically optimized for our application. This leaves room for further research and improvements, particularly in the decoder section of the architecture. Third, one potential limitation of our DA-TransUNet architecture is the risk of losing fine-grained details during the tokenization process, which occurs after the convolution and pooling operations in the encoder. This is particularly concerning for medical images with thin and complex structures, where preserving intricate details is crucial for accurate segmentation. Although our proposed integration of the Dual Attention (DA) module before the transformer in the encoder and within the skip connections helps mitigate this issue to some extent, as evidenced by the improved segmentation performance, this study acknowledges that there may still be room for further enhancement in capturing and retaining fine-grained information.

In addition, in the design of the Dual Attention (DA) module, this study considered the relative importance of channel attention and position attention for improving the model's performance. The channel attention focuses on capturing the inter-dependencies between different channels, while the position attention emphasizes the spatial relationships between different positions. Both of these attention mechanisms contribute to the model's ability to learn discriminative features. However, the question arises as to whether one of these attention mechanisms should be considered as predominant and the other as auxiliary.

To address this question, this study conducted experiments and analysis to determine the optimal configuration of channel and position attention within the DA module. The results of these experiments led to the introduction of the Mutual Inclusion mechanism in the subsequent chapter. The Mutual Inclusion mechanism aims to effectively integrate channel and position attention, allowing them to mutually enhance each other. By treating both attention mechanisms as equally important and facilitating their interaction, the Mutual Inclusion mechanism seeks to further improve the model's ability to capture fine-grained details and enhance the overall

segmentation performance.

The introduction of the Mutual Inclusion mechanism represents a promising direction for future research, as it explores the synergistic relationship between channel and position attention. By optimizing the integration of these attention mechanisms, this study aims to develop more advanced and effective architectures for medical image segmentation. The next chapter will delve into the details of the Mutual Inclusion mechanism and present the experimental results demonstrating its effectiveness.

3.5 Chapter Summary

In this chapter, this study presented DA-TransUNet, a novel architecture for accurate medical image segmentation that strategically integrates Dual Attention (DA) blocks with a Transformer-based U-Net architecture. The proposed model leverages the strengths of both the attention mechanisms and the transformers to enhance feature extraction and improve segmentation performance.

The key contributions of this chapter are threefold. Firstly, this study proposed the integration of Vision Transformer (ViT) and Dual Attention (DA) blocks in the encoder of the U-Net architecture, which enhances the model's ability to capture both global and local features crucial for medical image segmentation. Secondly, this study introduced an optimized DA block tailored for medical image segmentation and incorporated it into each skip connection layer, enabling effective filtering of irrelevant information and refining the transmitted features. Third, this study extensively validated the segmentation performance and generalizability of DA-TransUNet on five diverse medical image segmentation datasets, demonstrating its superiority over state-of-the-art methods.

The experimental results on the Synapse multiorgan segmentation dataset showed that DA-TransUNet outperformed the baseline TransUNet model, achieving an average Dice Similarity Coefficient (DSC) of 79.80% and a Hausdorff Distance (HD) of 23.48 mm. Furthermore, DA-TransUNet consistently outperformed other state-of-the-art models in five additional medical image segmentation datasets, including CVC-ClinicDB, chest X-ray masks and labels, ISIC 2018 lesion segmentation, Kvasir-SEG polyp segmentation, and Kvasir-Instrument segmentation.

Ablation studies provided valuable information on the individual contributions of the DA blocks to the encoder and skip connections. The results highlighted the importance of strategically integrating attention mechanisms for improved medical image segmentation performance. Statistical analysis

further validated the significance of the improvements achieved by DA-TransUNet over the baseline TransUNet model.

Despite its advantages, DA-TransUNet has some limitations, such as increased computational complexity and the potential loss of fine-grained details during the tokenization process. Future research directions may include optimizing the decoder architecture, exploring more efficient attention mechanisms, and developing strategies to better preserve fine-grained information in medical images with complex structures.

In conclusion, DA-TransUNet represents a significant advancement in medical image segmentation by leveraging the power of attention mechanisms and transformers. The proposed architecture has the potential to impact clinical decision-making and patient care by providing accurate and robust segmentation results across a wide range of medical imaging modalities.

Chapter 4

MIPC-Net: Mutual Inclusion of Position and Channel Features for Precise Boundary Segmentation

4.1 Motivation and Objectives

Medical image segmentation plays an essential role in quantifying diseases, assessing prognosis, and evaluating treatment outcomes. It describes crucial observations in images, such as the degree, size, and location of the lesions. However, manual segmentation by experienced professionals is both time-consuming and tedious [1]. Therefore, with the advance of deep learning technologies, automatic medical image segmentation has attracted growing research interest.

Existing medical image segmentation methods usually follow the practice of combining Convolutional Neural Networks (CNNs) with Vision Transformer modules under the U-Net structure [51, 53, 70]. For example, various U-Net variants have been proposed to improve medical image segmentation performance. ResUnet [5], Unet++ [18], and Unet3++ [19] introduced residual connections and complex skip connections, while Attention-Unet [6] integrated attention mechanisms into the U-Net architecture. TransUNet [9] and Swin-Unet [11] incorporated the Transformer and Swin-Transformer [12] modules, respectively, to capture global information. However, medical image segmentation differs from generic image segmentation tasks. In medical image segmentation, data is characterized by small sample sizes and the need for precise boundary delineation. Unlike generic image segmentation models, which are required to cover all details of the image, medical image segmentation demands special attention to abnormal regions and boundary details in organ or pathological images. Therefore, the features of the local image must be combined with the global features. To this end, attention

mechanisms focusing on both channel and position information need to be introduced into the research.

In recent research, there has been a trend towards incorporating both channel and position attention mechanisms into models. SA-UNet [28] and AA-TransUNet [14] incorporated spatial and channel attention, respectively, but did not make full use of the features of the image. TransUNet++ [10] and DS-TransUNet [13] integrated Transformers into skip connections but have limitations in overall architecture and feature integration. DA-TransUNet [2] merges position and channel attention but merely adapts a block of road segmentation, lacking custom feature extraction for medical images. These methods achieve better performance over previous medical image segmentation models. However, they primarily focus on the overall segmentation overlap rather than specifically enhancing the boundary details of the segmentation results. Moreover, when extracting features from the perspective of channel and position, these models only focus on repeated feature extraction, potentially disrupting the original information without considering how to restore the boundary details of the image.

Inspired by radiologists' working patterns, this paper proposes a simple and effective mutual inclusion mechanism for medical image segmentation. Instead of simply stacking transformer-related modules, this study introduces the Mutual Inclusion of Position and Channel Attention (MIPC) module, which enhances the focus on channel information when extracting position features and vice versa. Figure 4.1 illustrates the superiority of the proposed mutual inclusion of position and channel attention compared to existing attention mechanisms. This study proposes two pairs of channel and position combinations, each pair emphasizing either channel or position information while mutually including the other. This approach mimics the radiologists' working patterns, where mutual inclusion is practiced with varying emphasis. The experimental results demonstrate that this method effectively improves the model's ability to accurately segment image boundaries. Furthermore, this study focuses on the restoration of medical images by proposing the GL-MIPC-Skip-Connection. This connection introduces a Dual Attention mechanism to filter out invalid information while utilizing a global residual connection to restore the most effective information lost during the feature extraction process.

This study evaluated the proposed methods on three publicly accessible datasets: the Synapse dataset [57], the ISIC2018-Task dataset [59, 60], and the Segpc dataset [71]. In addition to the Dice coefficient (DSC) metrics, which deal with class imbalance problems, this study adopts the Hausdorff distance (HD) to analyze the quality of the segmentation results, as it is particularly convincing in evaluating boundary region segmentations. The

results show that the proposed method achieves state-of-the-art performance on both DSC and HD metrics. Notably, there was a 2.23mm reduction over competing models in the HD metric on the benchmark Synapse dataset, strongly evidencing the model's enhanced capability for precise image boundary segmentation. This finding also indicates that medical image segmentation benefits from the mechanism of mutual inclusion of position and channel attention.

The main contributions are as follows:

1) This paper proposes a novel model, MIPC-Net, which incorporates a Mutual Inclusion attention mechanism for position and channel information. This approach further improves the precision of boundary segmentation in medical images.
2) This paper introduces the GL-MIPC-Residue, a global residual connection that improves image restoration by enhancing the integration of the encoder and decoder.
3) Experiments demonstrate that the proposed components achieve consistent performance improvements. Furthermore, the model achieves state-of-the-art performance in all metrics in the public Synapse [57], ISIC2018-Task [59, 60], and Segpc [71] datasets.

The rest of this article is organized as follows. Section II reviews the related work of automatic medical image segmentation, and the description of the proposed MIPC-Net is given in Section III. In Section IV, comprehensive experiments and visualization analyzes are then conducted. Finally, Section V draws a conclusion to the whole work.

4.2 MIPC-Net Architecture

In the following section, this study introduces the MIPC-Net architecture, as depicted in Figure 4.2. This study begins by providing an overview of the overall structure. Subsequently, this study presents its key components in the following sequence: Mutual Inclusion of Position and Channel (Section 4.2.2), encoder (Section 4.2.3), GL-MIPC-Skip connections (Section 4.2.4) and decoder (Section 4.2.5).

4.2.1 Overview of MIPC-Net

Figure 4.2 illustrates the detailed configuration of the MIPC-Net model, which is a medical image segmentation model capable of capturing image-specific channel and position information and incorporates improved skip

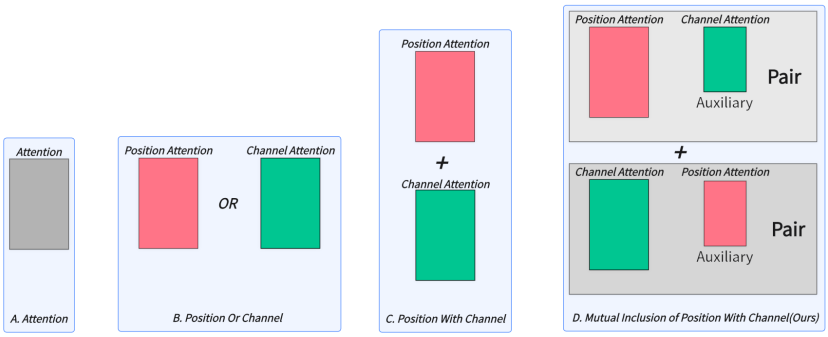

Figure 4.1: Comparison of attention mechanisms used in different medical image segmentation models: (a) only attention, (b) only channel or position attention, (c) integration of position and channel attention, and (d) Mutual nclusion of position and channel attention proposed in this work, which enhances the focus on channel information when extracting position features and vice versa"

connections.

The model consists of three main components: the encoder, the decoder, and the GL-MIPC-Skip connections. In particular, the encoder integrates the traditional convolutional neural network (CNN) and transformer mechanisms, while using MIPC-Block to enhance encoding capability (Section 4.2.3). The decoder relies on deconvolution to restore the features to the original feature map size (Section 4.2.5). GL-MIPC-Skip-Connections employ DA-Block to purify the features of skip connection transmission. Furthermore, they use the GL-MIPC-Residue to further enhance the integrity of the encoder and decoder (Section 4.2.4). MIPC-Net, made up of three integral components, exhibits superior image segmentation performance.

Given the constraints highlighted by traditional models, it is evident that while the conventional U-net architecture excels in capturing image features, it lacks effective methods for preserving and extracting global features.

On the other hand, Transformers exhibit remarkable proficiency in preserving and extracting global features through self-attention mechanisms [9]. However, they are inherently limited to unidirectional positional attention, overlooking the utilization of image-special position and channel. To address these limitations, this study has integrated the Mutual Inclusion of Position and Channel Block (MIPC-Block) and leveraged GL-MIPC-Skip-Connections to enhance the integrity of the encoder and decoder, thereby improving the performance of medical image segmentation.

In medical image segmentation tasks, current models usually use attention mechanisms to enhance the segmentation capabilities of the model.

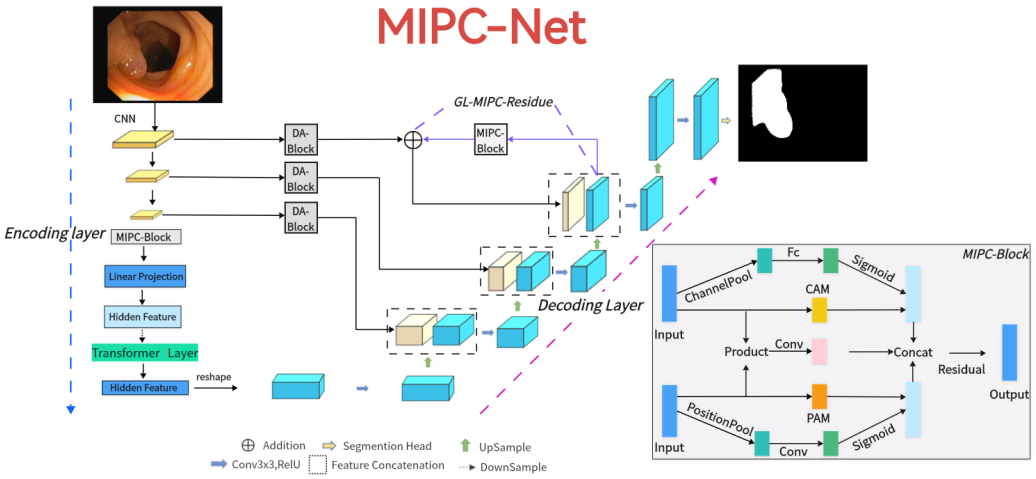

Figure 4.2: The illustration of the proposed MIPC-Net is depicted.

For example: TransUNet uses ViT, and Swin-Unet uses Swin-Transformer. These approaches do not adapt attention mechanisms to the specific features of the image and hence are unable to extract deep image-related information. To solve this problem, the proposed MIPC-Block enhances the segmentation capabilities of the model by leveraging image-specific features related to position and channel. It effectively combines these two features in a mutually inclusive manner to extract deeper image-related features, achieving subdivided extraction of image features and more fully mining features.

As illustrated in Figure 4.3, the MIPC-Block architecture seamlessly integrates image-specific channel and positional features, enriched by the application of residual concepts. The amalgamation of channel and positional features empowers the MIPC-Block with profound insight into the image, surpassing the capabilities of conventional attention modules.

The MIPC-Block architecture consists of three parts: PART A, PART B, and PART C. PART A and PART C serve as crucial feature extraction modules, ingeniously integrating both position and channel information of image features. The tight coupling of positional and channel information further enhances the feature extraction capability of the module. In Part A, the module undergoes a channel-wise average pooling layer (ChannelPool) to compress the feature map. Subsequently, it passes through fully connected layers to learn the correlations between different channels in the features. Following this, a sigmoid function is applied to constrain the values between 0 and 1, which yields channel correlations. Multiplying these correlations with the features obtained through the Position Attention

Module (PAM) results in information where the position is the primary focus and channels act as auxiliary. In contrast, in Part C, the features are first subjected to MaxPool and AvgPool operations (PositionPool) along the spatial dimensions. The resulting features from these two pooling operations are concatenated, and through fully connected layers, correlations between different spatial dimensions in the features are learned. Similarly to Part A, a sigmoid function constrains the values between 0 and 1. Multiplying these spatial correlations with the features obtained through the Channel Attention Module (CAM) produces information where channels are the main focus and spatial dimensions serve as auxiliary. Part B employs a residual approach to minimize the loss of valuable original information introduced by the convolution and attention modules.

Part A (Position-Dominant Extraction with Channel): As illustrated in Figure 4.3, ChannelPool facilitates the extraction of channel information from the input characteristics. Subsequently, a series of fully connected layers is employed to capture interchannel correlations, resulting in β^1. Currently, another set of input features is processed by the Position Attention Module (PAM) to extract position information features, resulting in β^2. Following sigmoid processing of β^1, it is multiplied element by element with β^2 to obtain β. In contrast to Part C, where channel-wise modulation is utilized for distributing feature maps from the spatial module, this process generates feature maps with spatial and channel emphasis.

$$\beta^1 = FC(ChannelPool(\text{Input})), \tag{4.1}$$

$$\beta^2 = PAM(\text{Input}), \tag{4.2}$$

$$\beta = Sigmoid(\beta^1) \cdot \beta^2, \tag{4.3}$$

PART B (Residual Part) : As shown in the figure, the inputs from Part A and Part B undergo a convolutional operation to obtain ω^1 and ω^2, respectively. Subsequently, the two are multiplied in element and then passed through another convolutional layer to yield ω. It extracts and refines the features of both inputs, thus refining the original features.

$$\omega^1 = Conv(Part A's \quad Input), \tag{4.4}$$

$$\omega^2 = Conv(Part C's \quad Input), \tag{4.5}$$

$$\omega = Conv(\omega^1 \cdot \omega^2) \tag{4.6}$$

PART C (Channel-Dominant Extraction with Position) : As shown in Figure 4.3, the input features undergo PositionPool along the

spatial dimension to effectively extract spatial information while eliminating noise and irrelevant details in the image. Subsequently, the feature maps are further processed by convolution to capture spatial correlations, resulting in α^1. Gleichzeitig, the Channel Attention Module (CAM) to extract channel characteristics, denoted α^2. The channel attention module is used to extract detailed channel features from the image. After sigmoid processing of α^1, it is multiplied in element by α^2 to obtain the output α. Unlike Part A, where the feature maps extracted by the spatial module are weighted by the channel attention module, effectively integrating image-specific spatial and channel features, generating feature maps with channel emphasis and spatial emphasis.

$$\alpha^1 = Conv(PositionPool(\text{Input})), \tag{4.7}$$

$$\alpha^2 = CAM(\text{Input}), \tag{4.8}$$

$$\alpha = Sigmoid(\alpha^1) \cdot \alpha^2, \tag{4.9}$$

Finally, the outputs of Parts A, B, and C are summed along the channel dimension, and then passed through a residual network (see Figure 4.2.2) to obtain the output.

$$\text{Output} = \text{Residual}(\alpha + \beta + \omega), \tag{4.10}$$

4.2.2 Mutual Inclusion of Position and Channel

The Mutual Inclusion of Position and Channel Block (MIPC-Block) mutually includes the image features' position and channel, capturing deeper features associated with image features compared to standard attention modules.

4.2.3 Encoder

As shown in Figure 4.2, the encoder consists of four key components: convolution blocks, MIPC-Block, an embedding layer, and transformer layers.

It is particularly significant that the MIPC-Block is introduced just before the transformer layers. The purpose is to subject the convolutional features to specialized image processing, enhancing the transformer's feature extraction capabilities with respect to the image's content. The Transformer architecture excels at capturing global information. Integrating the MIPC-Block enhances its ability to maintain and extract global features specifically

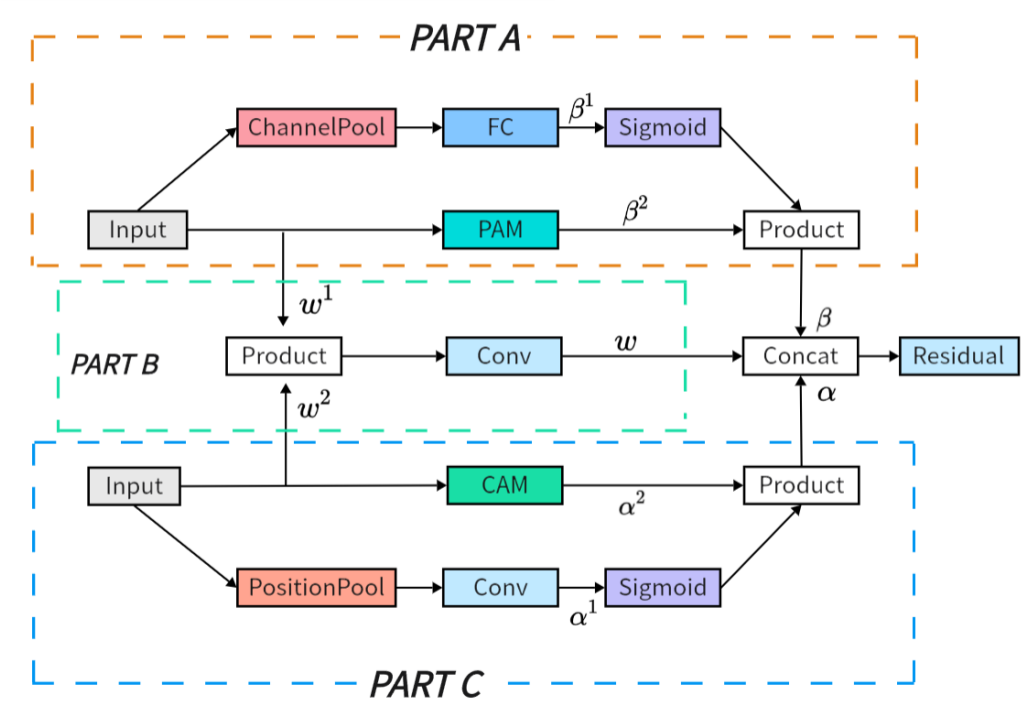

Figure 4.3: The proposed Position and Channel Mutual Inclusion Block (MIPC-Block) .

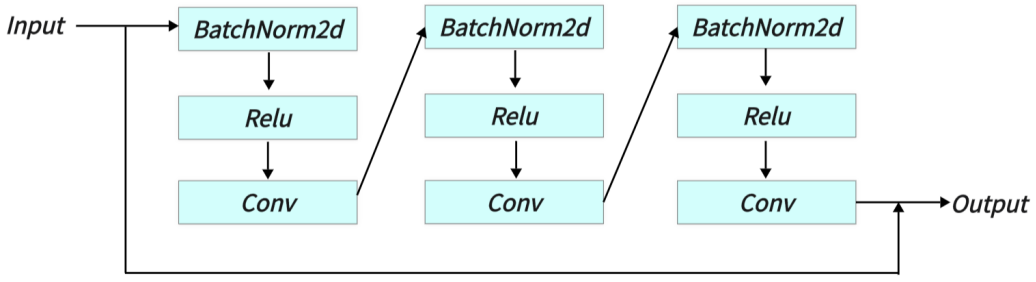

Figure 4.4: The specific structure of the last Residual module in MIPC-Block.

from images, enriching the Transformer's image-processing capabilities. This approach effectively combines image-specific channel and positional features with global features.

It begins with three convolutional blocks of the U-Net. Each block consists of a series of convolutions, normalization, and activation, designed

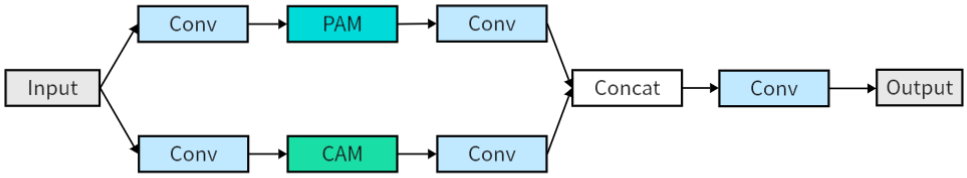

Figure 4.5: Architecture of Dual Attention Block (DA-Block).

to progressively refine the input features, halve their size, and double their dimensions, thereby achieving efficient feature extraction. The MIPC-Block then purifies these features, highlighting specific image details for a deeper analysis. An embedding layer adjusts the feature dimensions for transformer layers, which address CNN limitations by capturing global information. Finally, the transformer's output is recombined and directed through skip connections to the decoder, ensuring comprehensive information retention and enhancing segmentation performance in a streamlined process.

By incorporating convolutional neural networks, transformer architecture, and Mutual Inclusion of Position and Channel, the encoder configuration ultimately attains robust feature extraction capabilities, resulting in synergistic strength.

4.2.4 GL-MIPC-Skip-Connections

Within the framework of the U-shaped encoder-decoder architecture, skip connections are utilized to alleviate semantic discrepancies between encoder and decoder components. However, optimization of skip connections remains an area in need of improvement. Primarily, there exist challenges such as loss of feature fidelity during transmission and insufficient overall integrity between the encoder and decoder. To address these issues, this study employed two strategies: purifying the features transmitted via skip connections and augmenting skip connections with global information. These approaches facilitate the decoder in accurately restoring the original feature map, thereby significantly enhancing the model's segmentation capabilities. Here, this study calls the entire skip connection part GL-MIPC-SKip-Connections. It is divided into two parts: DA-SKip connections and GL-MIPC residue.

4.2.4.1 DA-Skip Connections

Analogously to conventional U-structured models [51] [7], the approach utilizes traditional skip connections to diminish the semantic disparity between the encoder and decoder. To further narrow this gap, this study has incorporated dual attention blocks (DA-Blocks) within the three skip connections, as illustrated in Figure 4.5. This enhancement stems from the observation that features conveyed through skip connections frequently harbor redundancies, which DA-Blocks are adept at filtering out, thereby refining the feature transmission process.

The integration of Dual Attention Blocks (DA-Blocks) into skip connections empowers the model to meticulously refine features relayed from the encoder, through the lens of image-specific positional and channel-based considerations. This process facilitates the extraction of more relevant information while minimizing redundancy. Such an enhancement not only bolsters the model's robustness, but also significantly reduces the likelihood of overfitting, thereby contributing to superior performance and enhanced generalization capabilities.

4.2.4.2 GL-MIPC-Residue

The distinction from other U-structured models lies in the sophisticated refinement of the decoder features and their strategic incorporation into the skip connections, as illustrated in Figure 4.2. This approach is motivated by the realization that, although encoder features are extensively leveraged via skip connections, decoder features often remain underexploited. By purifying the features of the decoder prior to their integration into skip connections, thus enhancing the restoration process of the original feature map, this study facilitate a deeper use of the features of the decoder.

Purifying features within the decoder, after three stages of upsampling, using Mutual Inclusion of Position and Channel (MIPC-Blocks), oriented specifically towards image-relevant channels and positions, significantly elevates the quality of information. Subsequent transmission of these enhanced features to skip connections, followed by their integration into the decoder, ensures the complete utilization of the decoder features. This methodology effectively minimizes redundancy between the encoder and the decoder, enriches the depth of the features, mitigates overfitting risks, and improves the model's capabilities in image segmentation and generalization.

4.2.5 Decoder

As depicted in Figure 4.2, the right-hand section of the diagram represents the decoder. The decoder's fundamental task is to leverage features sourced from the encoder and those transmitted via skip connections. Through processes including upsampling, it endeavors to accurately reconstruct the original feature map.

The decoder architecture is structured around three pivotal elements: feature fusion, the segmentation head, and a series of three upsampling convolution blocks. Initially, feature fusion operates by amalgamating feature maps received through skip connections with current feature maps, thereby equipping the decoder to accurately reconstitute the original feature map. Subsequently, the segmentation head undertakes the task of adjusting the final output feature map back to its original dimensions. The final element comprises three upsampling convolution blocks, methodically increasing the size of the input feature map at each stage to adeptly reinstate the image's resolution.

Due to the synergistic operation of these three components, the decoder showcases formidable decoding prowess. It adeptly harnesses features conveyed via skip connections, as well as those derived from intermediate layers, enabling a proficient reconstruction of the original feature map.

4.3 Experiment and Results

4.3.1 Datasets

The experiments are conducted on two distinct datasets: Synapse [57], ISIC 2018 [59, 60] and Segpc [71] for the following reasons:

Firstly, the Synapse dataset is among the most frequently utilized benchmark datasets in medical image segmentation, featuring segmentation tasks for eight different organs. This variety not only challenges, but also demonstrates the generalization capabilities of the model across diverse anatomical structures.

Secondly, the selection encompasses both a 3D multiclass segmentation challenge (Synapse) and a 2D single-class segmentation task (ISIC 2018, Segpc). This combination allows us to evaluate the model's segmentation abilities from different perspectives, effectively showcasing its versatility and robustness in handling both complex three-dimensional data and simpler two-dimensional images.

This strategic choice of datasets underscores the commitment to validat-

ing the model's performance across a range of segmentation tasks, highlighting its potential for widespread application in medical image analysis.

4.3.1.1 Synapse

The Synapse dataset comprises 30 CT scan images encompassing 8 abdominal organs, including the left kidney, right kidney, aorta, spleen, gallbladder, liver, pancreas, and stomach. In total, 3779 abdominal CT images enhanced with axial contrast were obtained. The in-plane resolution of these images varies from 0.54 x 0.54 mm^2 to 0.98 x 0.98 mm^2, while the slice thickness ranges from 2.5 mm to 5.0 mm.

4.3.1.2 ISIC-2018-Task

The dataset used in the 2018 ISIC Challenge addresses the challenges of skin diseases. It comprises a total of 2512 images, with a file format of JPG. The images of lesions were obtained using various dermatoscopic techniques from different anatomical sites (excluding mucous membranes and nails). These images are sourced from historical samples of patients undergoing skin cancer screening at multiple institutions. Each lesion image contains only a primary lesion.

4.3.1.3 Segpc

This challenge targets robust segmentation of cells and is the first stage in the construction of such tools for plasma cell cancers known as multiple myeloma (MM), a blood cancer. Provides images of normalized stained colors. The dataset contains a total of 298 images.

4.3.2 Implementation Settings

4.3.2.1 Baselines

In order to innovate in the field of medical image segmentation, this study conducted benchmark testing of the proposed model against a series of well-regarded baselines, including U-net, UNet++, Residual U-Net, Att-UNet, TransUNet, and MultiResUNet. U-net has been a foundational model in the medical image segmentation domain [51]. UNet++ enriches the skip connections [18]. Residual U-Net integrates a single residual module into the U-Net model [5], while MultiResUNet incorporates multiple residual modules [69]. Att-UNet utilizes attention mechanisms to improve the weight

of feature maps [6]. Finally, TransUNet integrates the Transformer architecture, establishing a new benchmark in segmentation accuracy [9]. Through comprehensive comparisons with these renowned baselines, the objective is to highlight the unique advantages and wide-ranging potential applications of the proposed model. In addition, this study benchmarked the model against advanced models. UCTransNet allocates attention modules in the traditional U-net model for skip connections [65], while MISSFormer moves attention module allocation into a transformer module-based U-shaped structure [72]. TransNorm integrates Transformer modules into the encoder and skips standard U-Net connections [66]. A novel transformer module was designed and a model named MT-UNet was constructed with it [32]. Swin-UNet further enhances segmentation by extensively applying Swin-transformer modules [11]. DA-TransUNet enhances the model segmentation capabilities by using image feature location contracts [2]. Through extensive comparisons with current state-of-the-art solutions, this study aim to showcase its outstanding segmentation performance.

4.3.2.2 Implementation Details

This study implemented MIPC-Net using the PyTorch framework and trained it on a single NVIDIA RTX 3090 GPU [68]. The Transformer module used in this study employs the pre-trained model "R50-ViT". The input resolution and patch size are set to 224x224 and 16, respectively. This study trained the model using the SGD optimizer, setting the learning rate to 0.01, the momentum to 0.9, and the weight decay to 1e-4. The default batch size was set to 24. The loss function employed for dataset is defined as follows:

$$\text{Loss} = \frac{1}{2} \times \text{Cross-Entropy Loss} + \frac{1}{2} \times \text{DiceLoss} \tag{4.11}$$

4.3.2.3 Model Evaluation

When evaluating the performance of MIPC-Net, this study utilizes a comprehensive set of metrics, including Intersection over Union (IoU), Dice Coefficient (DSC) and Hausdorff Distance (HD). These metrics are industry standards for computer vision and medical image segmentation and allow a multifaceted assessment of a model's accuracy, precision, and robustness.

AC(Accuracy): Accuracy is a widely used metric that assesses the overall correctness of a model's predictions. Calculate the proportion of correctly predicted samples over the total number of samples.Accuracy gives a general

idea of how well the model performs across all classes.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.12}$$

PR (Precision): Precision focuses on the accuracy of the positive predictions made by the model. Precision is the ratio of correctly predicted positive observations to the total predicted positives.High precision indicates that the model is good at not misclassifying negative instances as positive.

$$PR = \frac{TP}{TP + FP} \tag{4.13}$$

SP (Specificity): Specificity measures the accuracy of the negative predictions made by the model. Specificity is the ratio of correctly predicted negatives to the total predicted negatives. High specificity suggests that the model is effective at correctly identifying true negatives.

$$SP = \frac{TN}{TN + FP} \tag{4.14}$$

In summary, accuracy provides an overall view of model performance, precision emphasizes positive predictions' accuracy, and specificity assesses the accuracy of negative predictions.

IOU (Intersection over Union) is one of the commonly used indicators to evaluate the performance of computer vision tasks such as target detection, image segmentation, and instance segmentation. Measures how much the predicted area of the model overlaps with the actual target area, helping us to understand the accuracy and precision of the model. In image segmentation and instance segmentation tasks, IOU is used to assess the degree of overlap between predicted regions and ground-truth segmentation regions.

$$IOU = \frac{TP}{FP + TP + FN} \tag{4.15}$$

The Dice coefficient (also known as Srensen-Dice coefficient, F1 score, DSC) is a measure of model performance in image segmentation tasks and is particularly useful for dealing with class imbalance problems. Measures the degree of overlap between prediction results and ground-truth segmentation results and is particularly effective when dealing with object segmentation with unclear boundaries. The Dice coefficient is commonly used in image segmentation tasks as a measure of the accuracy of the model in the target area.

$$\text{Dice}(P, T) = \frac{|P_1 \cap T_1|}{|P_1| + |T_1|} \Leftrightarrow \text{Dice} = \frac{2|T \cap P|}{|F| + |P|} \tag{4.16}$$

Hausdorff distance (HD) is a distance metric that is used to measure the similarity between two sets and is often used to evaluate the performance of models in image segmentation tasks. It is particularly useful in the field of medical image segmentation, where it can quantify the difference between predicted and true segmentations, and it is particularly convincing in evaluating boundary region segmentations. The calculation of the Hausdorff distance captures the maximum difference between the true and predicted segmentation results.

$$H(A,B) = \max\left\{\max_{a\in A}\min_{b\in B}\|a-b\|, \max_{b\in B}\min_{a\in A}\|b-a\|\right\} \tag{4.17}$$

This study used Dice and HD in the Synapse dataset, used AC, PR, SP, Dice in the ISIC-2018-Task and Segpc datasets.

4.3.3 Comparison to the State-of-the-Art Methods

4.3.3.1 Synapse

To evaluate the performance of the proposed MIPC-Net model, this study conducted extensive experiments on the widely used Synapse multiorgan segmentation data set [57]. Using 12 state-of-the-art (SOTA) methods, including CNN-based and transformer-based approaches, such as U-Net [51], Res-Unet [5], TransUNet [9], U-Net++ [18], Att-Unet [6], TransNorm [66], UCTransNet [65], MultiResUNet [69], Swin-Unet [11], MT-UNet [32], and DA-TransUNet [2]. The experimental results are presented in Table 4.1.

As shown in Table 4.1, MIPC-Net achieves the highest mean Dice Similarity Coefficient (DSC) of 80.00% and the lowest average Hausdorff Distance (HD) of 19.32 mm among all the compared methods. This demonstrates the superior performance of MIPC-Net in both the overall segmentation accuracy and the boundary delineation precision. Compared to the popular transformer-based model TransUNet [9], MIPC-Net significantly improves the DSC by 2.52% and reduces the HD by 12.37 mm, highlighting the effectiveness of the proposed mutual inclusion mechanism and global integration strategy.

In addition, MIPC-Net consistently outperforms TransUNet in terms of DSC for the eight individual organs, with improvements ranging from 0.07% to 4.12%. In particular, MIPC-Net achieves substantial DSC improvements of 3.29%, 3.35%, 3.59%, 4.12%, and 3.93% for the gallbladder, right kidney, pancreas, spleen, and stomach, respectively. These organs are known to be particularly challenging to segment due to their variable shapes, sizes, and locations, as well as their low contrast with surrounding tissues. The

significant performance gains achieved by MIPC-Net demonstrate its strong capability in handling these difficult cases and accurately delineating organ boundaries.

Figure 4.6 provides a visual comparison of the DSC and HD values achieved by MIPC-Net and several other advanced models on the Synapse dataset. It is evident that MIPC-Net achieves the highest DSC and the lowest HD among all the compared models, further confirming its state-of-the-art performance in multi-organ segmentation.

To gain deeper insights into the boundary delineation performance of MIPC-Net, this study also evaluated the HD metric for each individual organ, as shown in Table 4.2. MIPC-Net achieves the lowest HD for five out of eight organs, including the aorta, gallbladder, right kidney, pancreas, and stomach. In particular, MIPC-Net significantly reduces HD by 6.31 mm and 2.73 mm for the aorta compared to TransUNet and DA-TransUNet, respectively. These results highlight the superior boundary segmentation capability of MIPC-Net, which can be attributed to the effective integration of position and channel information through the proposed mutual inclusion mechanism.

It should be noted that while MIPC-Net achieves state-of-the-art performance, its computational efficiency is comparable to that of TransUNet. The image segmentation time of MIPC-Net is 38.51 ms, only slightly higher than TransUNet's 33.58 ms. This indicates that the superior performance of MIPC-Net does not come at the cost of significantly increased computational overhead, making it a practical solution for real-world clinical applications.

Figure 4.7 presents a qualitative comparison of the segmentation results produced by TransUNet and MIPC-Net on the Synapse dataset. The regions highlighted by orange borders clearly demonstrate that MIPC-Net generates more accurate and precise segmentations compared to TransUNet, especially in challenging areas such as organ boundaries and small structures. The visual results further validate the effectiveness of the proposed approach in capturing fine-grained details and producing high-quality segmentation masks.

4.3.3.2 ISIC 2018-Task Dataset

To further validate the generalizability of MIPC-Net, this study performed experiments on the ISIC 2018 dataset [59, 60] for skin lesion segmentation. This dataset presents unique challenges, such as varying lesion sizes, shapes, and color variations.

Table 4.3.3.2 compares MIPC-Net with several state-of-the-art models on the ISIC 2018 dataset. MIPC-Net achieves the highest Accuracy

Table 4.1: The experimental results on the Synapse dataset include the average Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) for each organ, as well as the individual DSC for each organ.

		mDSC, mHD		DSC of a single organ							
Model	Year	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-net [51]	2015	76.85%	39.70	89.07	69.72	77.77	68.6	93.43	53.98	86.67	75.58
U-Net++ [18]	2018	76.91%	36.93	88.19	68.89	81.76	75.27	93.01	58.20	83.44	70.52
Residual U-Net [5]	2018	76.95%	38.44	87.06	66.05	83.43	76.83	93.99	51.86	85.25	70.13
Att-Unet [6]	2018	77.77%	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
MultiResUNet [69]	2020	77.42%	36.84	87.73	65.67	82.08	70.43	93.49	60.09	85.23	75.66
TransUNet [9]	2021	77.48%	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet [65]	2022	78.23%	26.75	84.25	64.65	82.35	77.65	94.36	58.18	84.74	79.66
TransNorm [66]	2022	78.40%	30.25	86.23	65.1	82.18	78.63	94.22	55.34	89.50	76.01
MT-UNet [32]	2022	78.59%	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
swin-unet [11]	2022	79.13%	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
DA-TransUNet [2]	2023	79.80%	23.48	86.54	65.27	81.70	80.45	94.57	61.62	88.53	79.73
MIPC-Net		**80.00%**	**19.32**	87.30	66.43	83.24	80.37	94.48	59.45	89.20	79.55

Table 4.2: The Hausdorff Distance (HD) for each organ in the Synapse dataset experimental results.

Model	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	14.94mm	15.81mm	59.92mm	45.76mm	37.86mm	17.34mm	43.33mm	18.56mm
swin-unet	8.64mm	27.98mm	41.83mm	34.00mm	22.17mm	12.43mm	**9.90mm**	15.45mm
DA-TransUNet	11.37mm	27.93mm	**30.76mm**	48.93mm	**20.26mm**	12.29mm	12.91mm	23.37mm
MIPC-Net	**8.63mm**	**15.74mm**	41.65mm	**27.12mm**	22.33mm	**11.58mm**	12.09mm	**15.39mm**

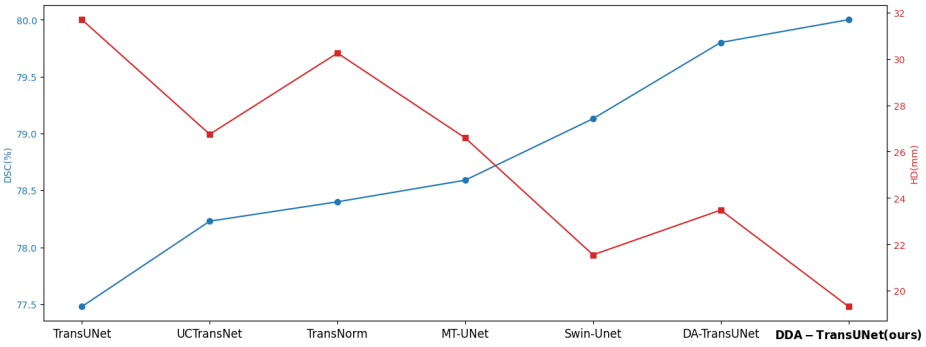

Figure 4.6: Line chart of DSC and HD values of several advanced models in the Synapse dataset

(AC) of 0.9560, Precision (PR) of 0.9279, and Specificity (SP) of 0.9831, demonstrating its superior performance in accurately segmenting skin lesions. Notably, MIPC-Net significantly outperforms the transformer-based model TransUNet, with improvements of 0.0108 in AC, 0.0453 in PR, 0.0178 in SP, and 0.0376 in Dice index. These improvements can be attributed to the effectiveness of the proposed mutual inclusion mechanism and global integration strategy in capturing both local and global contextual information.

Interestingly, while MIPC-Net achieves the highest AC, PR, and SP, its

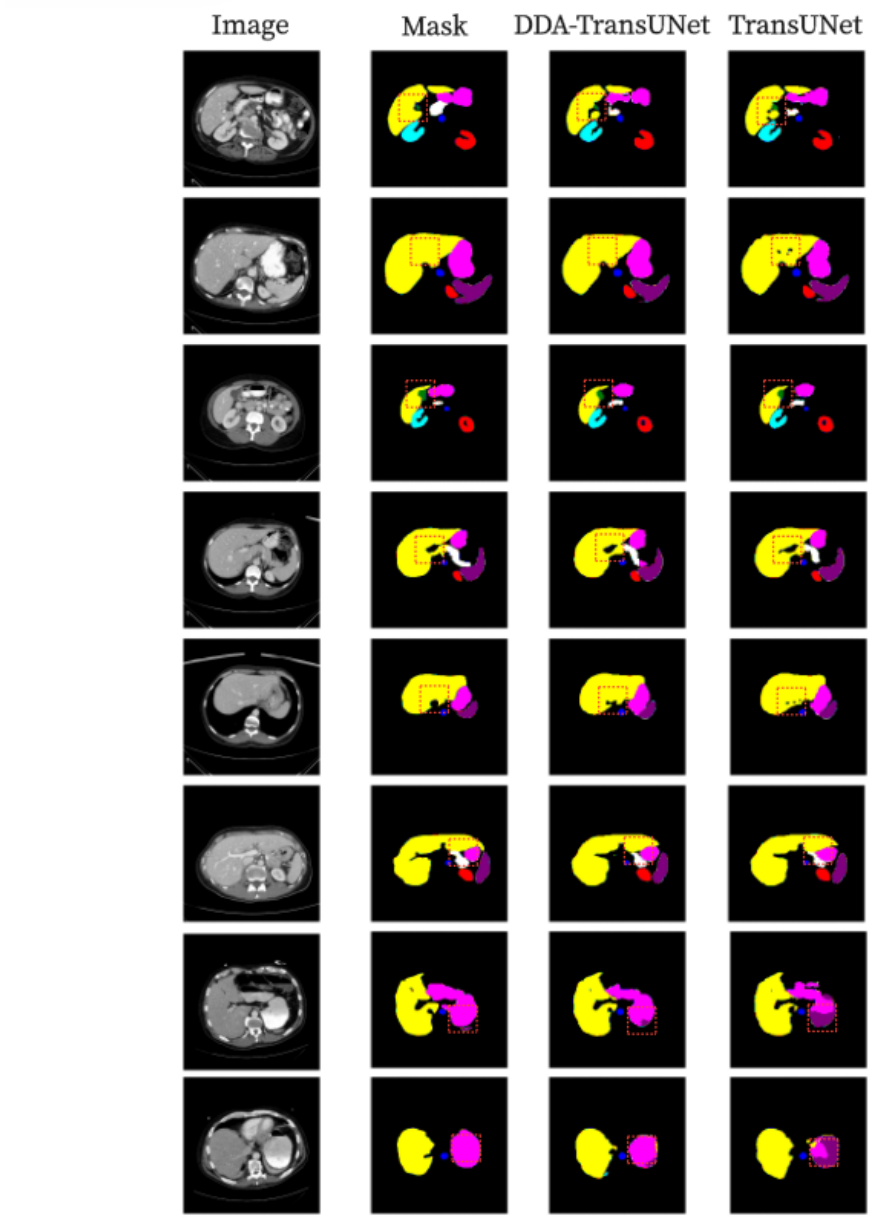

Figure 4.7: Segmentation results of TransUNet and MIPC-Net on the Synapse dataset.

Dice index of 0.8875 is slightly lower than that of UCTransNet (0.8898). This suggests a potential trade-off between precision and recall, which could be

Table 4.3: Experimental results on the ISIC2018-Task dataset

Method	AC	PR	SP	Dice
U-Net [51]	0.9446	0.8746	0.9671	0.8674
Att-UNet [6]	0.9516	0.9075	0.9766	0.8820
U-Net++ [18]	0.9517	0.9067	0.9764	0.8822
MultiResUNet [69]	0.9473	0.8765	0.9704	0.8694
Residual U-Net [5]	0.9468	0.8753	0.9688	0.8689
TransUNet [9]	0.9452	0.8823	0.9653	0.8499
UCTransNet [65]	0.9546	0.9100	0.9770	**0.8898**
MISSFormer [72]	0.9453	0.8964	0.9742	0.8657
MIPC-Net(ours)	**0.9560**	**0.9279**	**0.9831**	0.8875

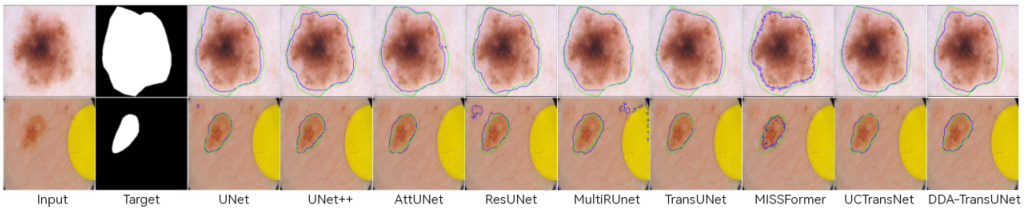

Figure 4.8: Segmentation results of TransUNet and MIPC-Net on the ISIC2018-Task dataset.

further investigated in future work.

Figure 4.8 qualitatively compares the TransUNet and MIPC-Net segmentation results on the ISIC 2018 dataset. MIPC-Net generates more precise and accurate segmentations, especially in challenging cases with irregular lesion boundaries and low contrast. The visual results further validate the superiority of the approach in capturing fine-grained details and producing high-quality segmentation masks for skin lesions.

4.3.3.3 Segpc Dataset

This study further assessed the performance of MIPC-Net on the Segpc dataset [71] for cell segmentation in microscopy images. This data set presents challenges such as overlapping cells, variable cell sizes and shapes, and low contrast between cells and background.

Table 4.3.3.3 compares MIPC-Net with state-of-the-art models on the Segpc dataset. MIPC-Net consistently outperforms all compared methods, achieving the highest Accuracy (AC) of 0.9817, Precision (PR) of 0.9079, Specificity (SP) of 0.9898, and Dice index of 0.8675. Compared to Tran-

Table 4.4: Experimental results on the Segpc dataset

Method	AC	PR	SP	Dice
Residual U-Net [5]	0.9733	0.8917	0.9871	0.8479
MultiResUNet [69]	0.9753	0.8391	0.9834	0.8613
TransUNet [9]	0.9671	0.8598	0.9882	0.8005
MISSFormer [72]	0.9663	0.8152	0.9823	0.8082
DA-TransUNet [2]	0.9713	0.8789	0.9845	0.8366
MIPC-Net(ours)	**0.9817**	**0.9079**	**0.9898**	**0.8675**

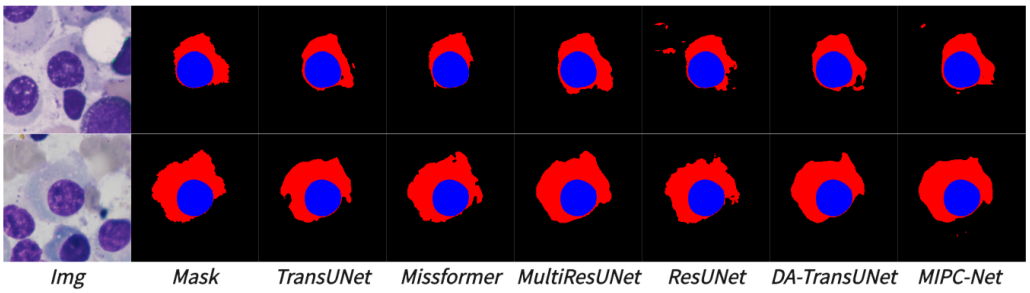

Figure 4.9: Segmentation results of TransUNet and MIPC-Net on the Segpc dataset.

sUNet, MIPC-Net significantly improves performance across all metrics, with improvements of 0.0146 in AC, 0.0481 in PR, 0.0016 in SP, and 0.067 in Dice index. These substantial improvements demonstrate the effectiveness of the approach in accurately separating overlapping cells and dealing with low contrast.

In particular, MIPC-Net achieves a significantly higher Dice index (0.8675) compared to all other methods, indicating a good balance between precision and recall when segmenting cells, which is crucial for accurate cell analysis and quantification.

Figure 4.9 visually compares the segmentation results of TransUNet and MIPC-Net on the Segpc dataset. MIPC-Net generates more accurate and precise segmentations, successfully separating individual cells and capturing their fine boundaries, even in dense cell clusters.

The strong performance of MIPC-Net on the ISIC 2018 and Segpc datasets, along with its state-of-the-art results on the Synapse dataset, highlights the versatility and generalizability of the approach across different medical image segmentation tasks and modalities.

4.3.4 Ablation Study

To gain a deeper understanding of the effectiveness of key components in the proposed MIPC-Net model, this study conducted a comprehensive ablation study on the Synapse data set. The study focused on three main aspects: the effects of mutual inclusion of position and channel, the impact of different configurations within the MIPC-Block, and the influence of the GL-MIPC-Residue in skip connections.

4.3.4.1 The effects of Mutual Inclusion of Position and Channel

Table 4.5: Effects of Mutual Inclusion of Position and Channel

	Mutual Inclusion	DSC↑	HD↓
PC-Net		79.09	23.34
MIPC-Net	√	**80.00**	**19.32**

As shown in Table 4.5, MIPC-Net, which incorporates the mutual inclusion mechanism, outperforms PC-Net by 0.91% in terms of DSC and achieves a reduction of 4.02mm in HD. This improvement can be attributed to the effective integration of position and channel information through the mutual inclusion mechanism. By allowing the position and channel attention modules to interact and mutually guide each other, MIPC-Net is able to capture more comprehensive and discriminative features, leading to more accurate and precise segmentations. In contrast, simply using position and channel information independently, as in PC-Net, fails to fully exploit the potential synergies between these two types of information, resulting in suboptimal performance.

4.3.4.2 The effects of how to mix MIPC-Block internal mechanisms

Table 4.6: Effects of how to mix MIPC-Block internal mechanisms

	Part.A Primary	Part.A Auxiliary	Part.C Primary	Part.A Auxiliary	DSC↑	HD↓
MIPC-Net	PAM	ChannelPool	CAM	PositionPool	**80.00**	**19.32**
MIPC-Net	PAM	ChannelPoll	PositionPool	CAM	78.87	21.55
MIPC-Net	ChannelPool	PAM	CAM	PositionPool	79.10	26.38
MIPC-Net	ChannelPool	PAM	PositionPool	CAM	79.11	24.27

Table 4.6 presents the results of different configurations within the MIPC-Block. The optimal configuration, where position attention (PAM) is used as the primary focus and channel attention (ChannelPool) as the auxiliary focus in Part A, and channel attention (CAM) is used as the primary

focus and position attention (PositionPool) as the auxiliary focus in Part C, achieves the best performance with a DSC of 80.00% and an HD of 19.32mm. This suggests that a balance between position and channel attention is crucial for achieving the best segmentation results. By employing different primary attention modules in Parts A and C, MIPC-Block is able to capture complementary information from both position and channel perspectives, leading to more comprehensive feature extraction. Furthermore, the results demonstrate that the use of PAM and CAM as primary attention modules consistently outperforms using ChannelPool and PositionPool as the primary modules, indicating that the self-attention mechanisms employed in PAM and CAM are more effective in capturing long-range dependencies and global contextual information.

4.3.4.3 The effect of the GL-MIPC-Residue in skip connections

Table 4.7: Effects of the GL-MIPC-Residue in skip connections

	GL-MIPC-Residue						
	1st	2nd	3rd	DA-Skip-Connections	Encoder with MIPC	DSC↑	HD↓
MIPC-Net				√	√	79.28	25.27
MIPC-Net	√			√	√	**80.00**	**19.32**
MIPC-Net		√		√	√	79.90	21.82
MIPC-Net			√	√	√	78.64	27.78
MIPC-Net	√	√	√	√	√	78.25	28.06
MIPC-Net						77.48	31.69

Table 4.7 shows the impact of the GL-MIPC-Residue module on the overall performance of MIPC-Net. Adding the GL-MIPC-Residue module to the first skip connection layer alone achieves the best performance, with a DSC of 80.00% and an HD of 19.32mm, outperforming the baseline MIPC-Net without any GL-MIPC-Residue by 0.72% in terms of DSC and reducing HD by 5.95mm. This suggests that the GL-MIPC-Residue module is most effective when applied to the shallower skip connection layers, particularly the first layer, as it captures more low-level and spatial information crucial for accurate boundary delineation. The GL-MIPC-Residue module provides a direct path for the propagation of high-resolution spatial information from the encoder to the decoder, helping to preserve fine-grained details and improve localization accuracy. However, applying the GL-MIPC-Residue module to all skip connection layers leads to a significant performance drop, indicating that excessive use of the module can be counterproductive.

In conclusion, the ablation study demonstrates the importance of the mutual inclusion mechanism, the careful design of attention mechanisms

within the MIPC-Block, and the strategic placement of the GL-MIPC-Residue module in skip connections. These components work together to capture comprehensive and discriminative features, leading to improved segmentation accuracy and precise boundary delineation in medical images.

4.3.5 Discussion

In this chapter, this study found that Mutual Inclusion of Image-specific Channels and Positions can provide significant assistance for Medical Image Segmentation Tasks. The proposed MIPC-Block, based on the Mutual Inclusion mechanism, combined with GL-MIPC-Residue, further enhances the overall integration of the encoder and decoder. The proposition has been validated through experiments on datasets, with the HD metric showing an improvement of 2.23mm compared to competing models on the Synapse dataset, demonstrating strong boundary segmentation capabilities.

Analyzing the ablation experiments validates the effectiveness of the proposed MIPC Block and GI-MIPC-Residue. Firstly, according to the experimental results presented in Tables 4.5 and 4.6, this study concluded that mutual inclusion of image feature positions and channels yields better performance compared to simple usage. Furthermore, as demonstrated by the results in Table 4.7, the GL-MIPC-Residue module improves the overall integrity of the encoder-decoder. This study concludes that reducing the loss of effective features is of paramount importance when exploring features in depth.

Despite these advantages, the model has some limitations. Firstly, the introduction of MIPC-Block and DA-Blocks leads to an increase in computational complexity. This added cost may pose a barrier for real-time or resource-constrained applications. Furthermore, this approach combines feature positions and channels attention with the Vision Transformer in a parallel manner, without achieving deep integration between them, indicating potential areas for further research and enhancement.

4.4 Chapter Summary

In this chapter, this study proposed MIPC-Net, a novel medical image segmentation model that introduces the Mutual Inclusion of Position and Channel (MIPC) attention mechanism and the GL-MIPC-Residue module for precise boundary segmentation. The MIPC-Block effectively captures image-specific features by mutually including position and channel information, enabling the model to extract more comprehensive and discriminative

features. The GL-MIPC-Residue module, strategically integrated into the skip connections, enhances the overall integration of the encoder and decoder, facilitating the preservation of fine-grained details and improving localization accuracy.

The effectiveness of MIPC-Net was extensively evaluated on three publicly available datasets: Synapse, ISIC 2018, and Segpc. The proposed model achieved state-of-the-art performance across all datasets, outperforming several well-established baselines and advanced models. Notably, MIPC-Net demonstrated a significant improvement of 2.23mm in the Hausdorff Distance (HD) metric on the Synapse dataset compared to competing models, highlighting its strong boundary segmentation capabilities. The model also exhibited superior performance in terms of Accuracy, Precision, Specificity, and Dice index on the ISIC 2018 and Segpc datasets, further validating its generalizability across different medical image segmentation tasks and modalities.

The ablation study provided valuable insights into the contributions of key components in MIPC-Net. The results confirmed the importance of the mutual inclusion mechanism in capturing complementary information from both position and channel perspectives, leading to more accurate and precise segmentations. The study also highlighted the optimal configuration of attention mechanisms within the MIPC-Block and the strategic placement of the GL-MIPC-Residue module in skip connections for the best performance.

Despite its advantages, MIPC-Net has some limitations. The introduction of MIPC-Block and DA-Blocks increases the computational complexity, which may be a concern for real-time or resource-constrained applications. Additionally, the current approach combines feature positions and channels attention with the Vision Transformer in a parallel manner, leaving room for further research on achieving deeper integration between these components.

In conclusion, MIPC-Net represents a significant advancement in medical image segmentation, leveraging the power of mutual inclusion attention mechanisms and global integration strategies for precise boundary delineation. The proposed model has the potential to greatly benefit clinical decision-making and patient care by providing accurate and reliable segmentation results across a wide range of medical imaging modalities. Future research directions may include exploring more efficient attention mechanisms, investigating deeper integration of position and channel attention with transformers, and adapting the model to additional medical image segmentation tasks and datasets.

Chapter 5

FKD-Med: Federated Learning and Knowledge Distillation for Privacy-Preserving and Efficient Medical Image Segmentation

5.1 Motivation and Objectives

Within the healthcare domain, the field of medical image segmentation has experienced a paradigm shift due to the advent of advanced deep learning techniques. Federated Learning (FL) in medical image segmentation allows institutions to collectively enhance models while protecting patient data privacy. However, the problem arises from the need to share sensitive medical data between institutions while ensuring efficient processing. Traditional segmentation methods struggle to balance data privacy with computational and communication efficiency. Enhancing communication efficiency in the context of FL has substantial practical value. It enables the inclusion of more extensive medical data from a broader range of hospitals for training, significantly expanding the scope and depth of medical research and patient care. This advancement is pivotal for the development of more accurate and comprehensive medical analysis tools, ultimately benefiting healthcare outcomes worldwide. In recent years, deep learning models in medical scenarios have increasingly incorporated larger parameter volumes, significantly intensifying the demand for efficient communication in FL applications [73–75]. Therefore, this challenge has received widespread attention, prompting extensive research efforts to address these communication inefficiencies in the training of medical models.

The study of medical image segmentation presents challenges due to the scarcity of high-quality medical imaging data, particularly in contexts involving user privacy concerns. Consequently, extensive research has been dedicated to improving segmentation model structures in the past

few decades [76].To address this bottleneck, the focus has largely been on innovation of neural network architectures. The U-net model [4], introduced in 2015, employed an encoder-decoder framework to amplify the precision of segmentation despite the constraints of limited datasets. Following this, subsequent studies such as Att-UNet [77], ResUNet [5], MLDA-Unet [78], TransUNet [9], and FHI-Unet [79] built on this foundational work. The primary objective of these advances has been to refine deep learning architectures with the ultimate goal of improving segmentation accuracy. However, these developments have overlooked the vital issue of increasing the pool of available training samples.

To address these limitations, FL stands out as a potent approach. Facilitates the amalgamation of data between various hospitals, ensuring the adherence to stringent privacy norms [35]. Although studies like [35] and [38] highlight the advantages of FL through approaches such as the decentralized MQTT framework for brain tumor segmentation and the label-agnostic FedMix method for diverse medical image segmentation, they often overlook the critical need for communication efficiency optimization in models with large parameters. However, the increasing size of the model and the associated increase in communication costs in FL restrict its application, limiting the inclusion of data to only a select number of medical institutions. Moreover, to facilitate the inclusion of more nodes in Federated Learning (FL), several studies have implemented Knowledge Distillation (KD) to improve communication efficiency. MetaFed, with its cyclic KD, extended FL to multiple federations, improving precision and reducing communication costs [47]. FedX introduced unsupervised learning with dual-sided KD, increasing performance in unsupervised algorithms [80]. Furthermore, some research combined KD with a federated UNet for land use classification, achieving notable improvements in model compression and accuracy [81]. However, there is a notable gap in the development of framework tools for the integration of Federated Learning (FL) and Knowledge Distillation (KD) specifically tailored to medical image segmentation. The clinical scenarios of FKD-Med underscore its practical relevance, showcasing the potential for application in diverse healthcare settings [82] [83].

Integrating FL with KD offers a potent mechanism for the reduction of model parameters, thus reducing both computational and communication costs in FL [84]. In Fig. 5.1, a schematic diagram serves as a representative example to elucidate the method employed in this study, specifically the integration of FL and KD for medical image segmentation. Without FL, the computation would be limited to the data set of a single hospital. FL allows the incorporation of Hospital-1, Hospital-2, and Hospital-3 datasets into the computation by interacting with the central server, effectively tripling the

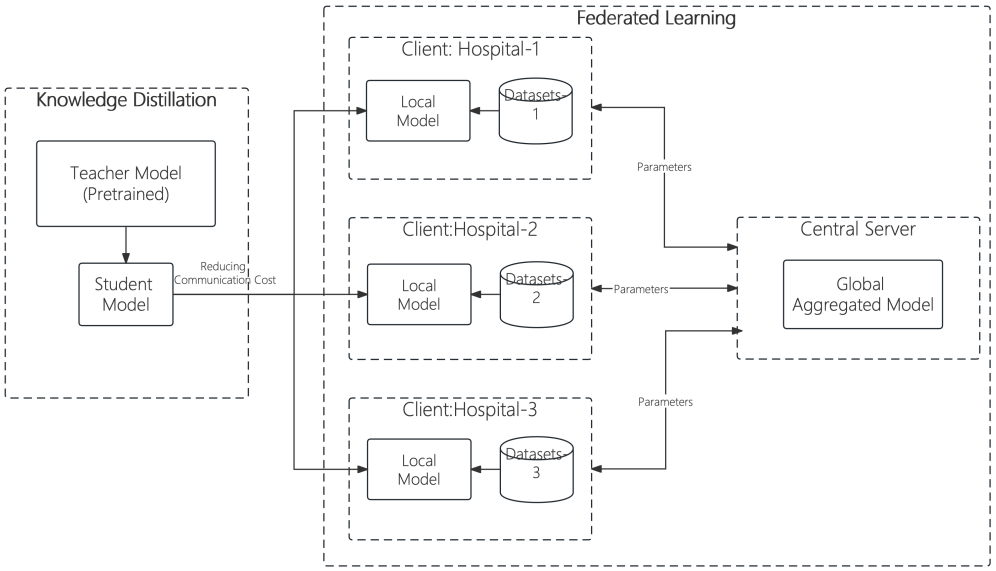

Figure 5.1: Toy example demonstrating the key principles of FKD-Med. The illustration simplifies the complex architecture into essential components, highlighting the interaction between FL and KD processes. This serves as a conceptual guide for understanding the integration of data aggregation and model optimization in FKD-Med.

total data volume for computation. Conversely, the left segment of Fig. 5.1 illustrates the KD mechanism, wherein a pretrained teacher model guides the learning process of a more compact student model. Through the use of this lightweight student model, each hospital only needs to exchange the parameters of its local student model with the central server, significantly improving communication efficiency. In conclusion, by employing both methods, this approach ensures that while training data volume is augmented through FL, communication efficiency and training speed are simultaneously optimized.

This study introduces an open-source, modular framework, initially engineered for the specialized requirements of medical image segmentation. This framework is designed with the versatility to be adapted for a broad spectrum of computational tasks in medicine, including, but not limited to, diagnostic analytics, treatment planning, and drug discovery. This framework seamlessly integrates FL and KD. In the healthcare setting, where the need for large training datasets must be balanced with data privacy

concerns, the framework relies on FL to meet both objectives effectively. Gleichzeitig, this study leverages KD techniques to increase training efficiency and overall model performance. The empirical validation is centered on medical image segmentation, utilizing datasets of CVC-ClinicDB [85] and Chest Xray [86] [87]. Here, pre-trained TransUnet [9] and ResUNet [5] act as teacher models, guiding the streamlined Tiny-Unet student model. In the experiments, the parameters of the student model were reduced to the 1 / 127 and 1/1027 fractions of the teacher models, resulting in an accuracy improvement of 0.25%, 0.43%, 1.35%, and 1.46%, respectively, compared to the scenario without KD. This not only substantiates the effectiveness of the framework in the realm of medical image segmentation, but also underscores its potential applicability to other medical computational areas, such as diagnostic analytics and treatment planning.

In this work, the contributions are as follows:

1. **Open-source Adaptable Framework:** The open-source framework, FKD-Med, offers versatility for a wide range of medical applications, extending beyond simple image segmentation.
2. **Pioneering Application of FL & KD for Medical Image Segmentation:** The work pioneers the application of FKD-Med for Medical Image Segmentation, merging FL with KD to cut communication costs in deep model training,
3. **Effective Reduction of Computation Costs and Protecting Privacy:** The framework significantly lowers computation and communication costs and preserves data privacy, compressing model parameters by factors of 127 and 1027 without sacrificing accuracy. This study validates this through experiments on two datasets.

The rest of the paper is organized as follows. Section II reviews the related work of FL and KD in the medical image segmentation task. The description of the proposed framework FKD-Med is given in Section III. The case study of two data sets and the experimental analysis are conducted in Section IV. Finally, Section V concludes the whole work.

5.2 The Proposed FKD-Med Framework

5.2.1 An Overview of The Framework

The primary novel contribution of the work is the FKD-Med framework, a unique fusion of FL and KD specifically designed for medical image segmentation. Unlike existing methods that apply either FL or KD, FKD-

Med synergistically combines these two techniques to address the challenges of limited training samples and high communication overhead. FKD-Med is a novel, open-source framework designed specifically for medical image analysis tasks, including segmentation and other computational processes related to medical data. It uniquely combines the principles of FL and KD, leveraging the strengths of both to provide a robust and efficient solution. To the best of knowledge, FKD-Med is the first framework that integrates Federated Learning and Knowledge Distillation for medical image segmentation.

The framework is equipped with a variety of U-Net-like models and loss functions, allowing for customization and flexibility based on the specific requirements of the task at hand. The primary aim of FKD-Med is to facilitate privacy-preserving, efficient and high-performing medical image segmentation, addressing some of the key challenges in the field.

The unique combination of FL and KD opens new vistas in clinical applications. Its potential extends to remote patient monitoring and telemedicine, where efficient data handling and preservation of privacy are crucial. In addition, FKD-Med paves the way for collaborative research in multiple healthcare settings, fostering a more inclusive and comprehensive approach to medical research. By allowing the amalgamation of diverse data sets while ensuring data privacy, FKD-Med stands as a cornerstone in the advancement of medical informatics and patient care.

The FKD-Med framework, as depicted in Fig 5.2 and Fig 5.3, is structured around three core components, making it a dynamic, efficient and versatile tool for medical image segmentation. The first component, the FL component, ensures data privacy through local data processing and collaborative model training. The second component, KD, boosts efficiency and reduces the interhospital communication costs that come with FL, by transferring complex models to simpler ones. Lastly, the U-Net-like model library and the loss function library offer a range of customizable models and selection of loss functions, respectively, suitable for various tasks. Together, these elements position FKD-Med as a robust and versatile tool, not limited to medical image segmentation. It is equally adept at handling a wide range of other medical applications, such as diagnostic decision support and predictive modeling for treatment outcomes.

5.2.2 Federated Learning in FKD-Med

The first key component is the FL component. The necessity of FL arises from the need to train machine learning models on distributed datasets, especially when data privacy and security are of paramount concern. In the medical field, patient data is often spread across different hospitals

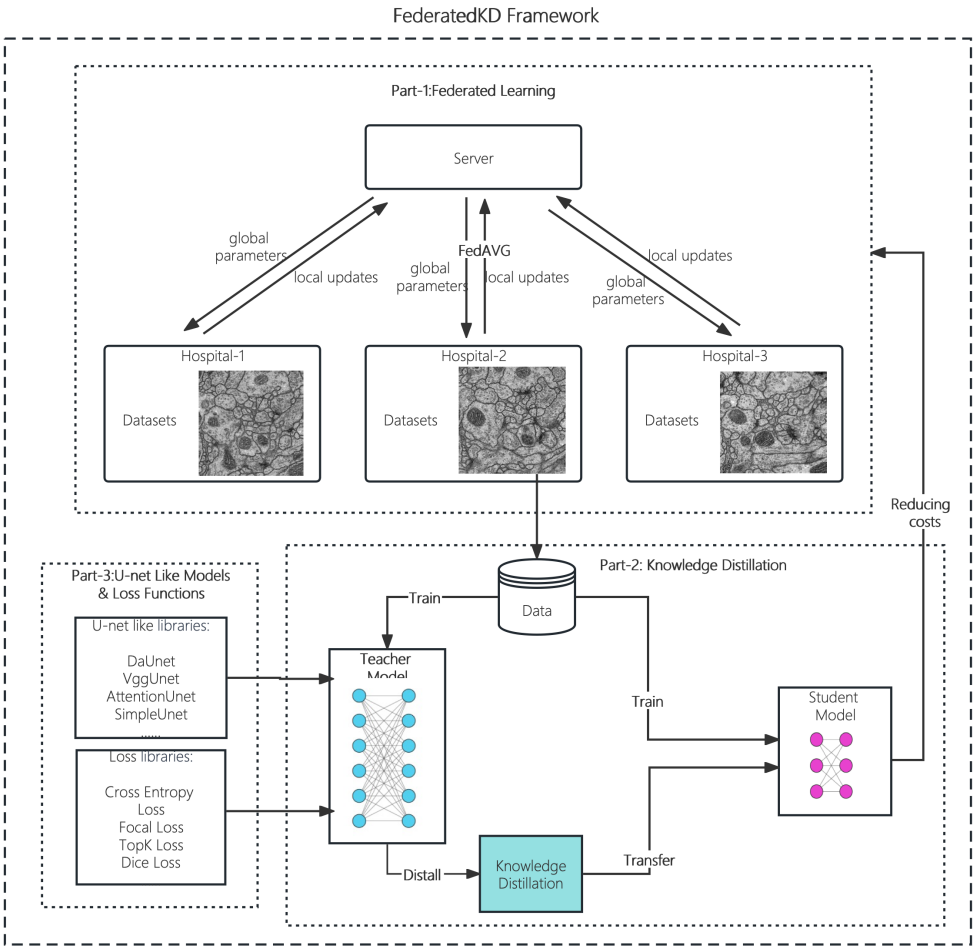

Figure 5.2: Schematic representation of the FKD-Med framework.

and institutions, and sharing raw patient data is often restricted due to privacy laws and regulations. FL enables us to leverage this distributed data for model training without compromising patient privacy. Although existing frameworks apply FL for data privacy, none have effectively reduced interhospital communication costs in the manner FKD-Med does.

FL is engineered to handle the distributed machine learning process, enabling data to be processed at local nodes, effectively safeguarding data privacy. To illustrate, consider a scenario where each of the three hospitals has a similar type of medical image. For example, in Fig. 5.2, the FL component orchestrates a collaborative training process between these three

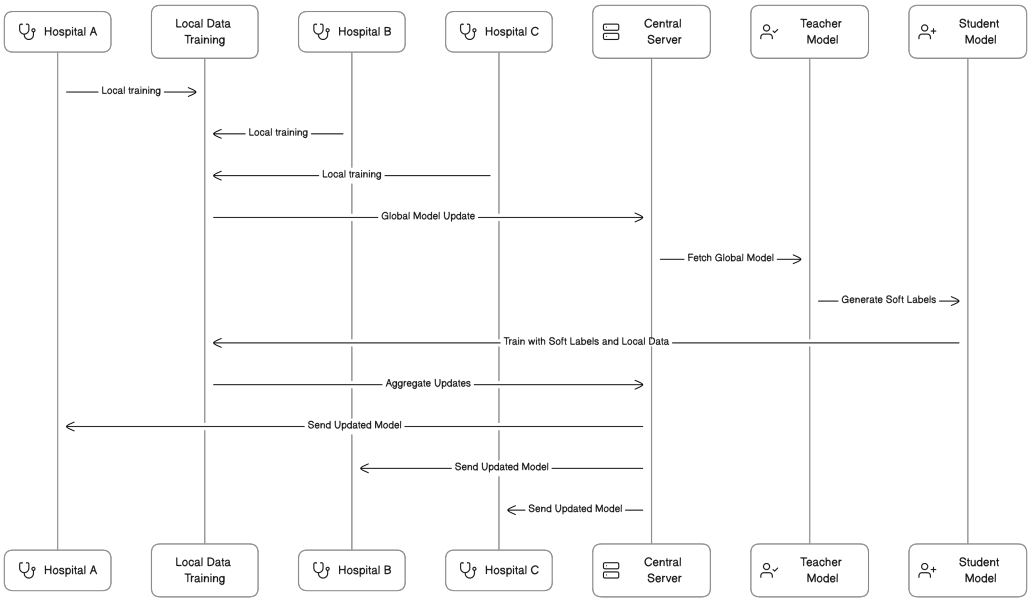

Figure 5.3: Swimlane Diagram of Modules Interaction in FKD-Med.

hospitals using the federated averaging (FedAvg) method [88]. This method allows each hospital to train models on their local datasets and then combines these locally refined models to form a comprehensive global model. This approach not only ensures the confidentiality of each hospital's data, but also capitalizes on the shared knowledge across all participating entities, thereby enhancing the overall performance of the model.

In the FKD-Med framework, FL is implemented using the federated averaging algorithm (FedAvg). This algorithm allows for the training of models on local datasets and the subsequent aggregation of these locally updated models to form a global model. The FedAvg algorithm is a key part of the FL component in the FKD-Med framework. Each participating institution (or 'client') trains the model on its local data and only shares the model parameters or updates with a central server. The server then aggregates these updates to improve the global model. This process repeats over multiple rounds until model performance meets the desired criteria. In this way, the raw data never leaves the local institutions, thus preserving data privacy. It begins with the weights of the global model as input. For each training round, the algorithm performs a local update on each client with the current global model weights and stores the local model weights. After training the local models, the global model weights are updated by taking the average of the local model weights. This process is repeated for

a certain number of rounds. The final global model weights are returned after all rounds of training are completed. This implementation of FL allows for efficient training of models on distributed data while reducing the communication overhead between different nodes, making the framework more efficient and cost-effective for medical image segmentation tasks.

In the FKD-Med framework, the implementation of FL is facilitated by the use of the Flower framework. Flower is a flexible, friendly, and fast machine learning framework for FL [89]. Provides a robust and efficient infrastructure to build and execute FL experiments. In the context of FKD-Med, Flower allows for efficient and scalable execution of the federated averaging algorithm (FedAvg). It enables the training of U-Net-like models on local datasets and the aggregation of locally updated models to form a global model. This approach ensures data privacy, reduces communication overhead, and enhances the overall performance of the model.

In the FKD-Med framework, this study applies FL to medical image segmentation. Each participating hospital trains the segmentation model locally using its own data. The model parameters are then shared with the central server, where they are aggregated to update the global model. This FL approach allows us to leverage a large amount of diverse data for model training while ensuring patient data privacy.

5.2.3 Knowledge Distillation in FKD-Med

The second integral component of the FKD-Med framework is the KD component, depicted in Fig. 5.4. Unlike traditional KD methods, FKD-Med's approach is uniquely designed to operate within federated environments and effectively reduce communication costs associated with FL. This distinguishes FKD-Med from existing methodologies.

FL trains models across decentralized devices, ensuring data privacy. In this setup, KD compresses bulky models into smaller, more efficient ones, addressing the rising communication overhead caused by complex model parameters, which hampers scalability and efficiency. This bottleneck, significant in transmitting extensive parameters over networks, is pivotal to improving FL's real-world viability, especially where communication resources are scarce. FKD-Med targets this by refining model size and performance balance, enhancing FL's practicality. In FKD-Med, knowledge is transferred from a complex 'teacher' model to a simpler 'student' model, slashing computational demands while maintaining performance. This strategy notably cuts communication costs among hospitals by minimizing data transmission needs, thereby boosting the framework's efficiency and cost-effectiveness.

As illustrated in Fig 5.4, the FKD-Med framework uniquely and effectively implements KD through a two-model process.

In the FKD-Med framework, the student model is intricately designed to handle medical images, producing dual outcomes for a comprehensive learning experience. The first, termed soft label, emerges from the final softmax layer of the model. This output is meticulously refined by aligning it with the soft labels from a pre-trained teacher model, ensuring a nuanced understanding of the data. The second outcome, known as the hard-label, is honed through direct comparisons with actual ground-truth labels, establishing a concrete benchmark for accuracy. The essence of this dual output mechanism lies in the integral role of the teacher model in refining both outputs of the student models. It meticulously guides the fine-tuning of the soft and hard labels, embodying a more detailed explanation of the distillation process tailored for medical image segmentation. This approach is strategically optimized not only to increase precision but also to minimize computational load and communication demands, illustrating a sophisticated balance between efficiency and effectiveness in model training.

The teacher model, known for its complexity and the large number of parameters, is adept at achieving high accuracy. It is often selected from models like Unet++, TransUnet, Swin-Unet, etc., that have demonstrated state-of-the-art performance in medical image segmentation tasks and is pre-trained on the entire training dataset with its parameters saved for future use. In contrast, the student model is a more primitive tiny U-net configuration, containing convolutional and deconvolutional layers and characterized by fewer parameters. Comprising two layers of upsampling and two layers of downsampling, the design of the student model facilitates a streamlined but effective approach [4]. Together, these models contribute to FKD-Med's robust capability in medical image analysis tasks.

In the KD process, both the teacher and student models receive the same input data. The teacher model processes the input data and outputs them to a softmax function with a temperature parameter $Temperature$ to penalize the loss as in [90] shown in Equation (5.1). The softmax function is applied to each element $\mathbf{x}_i$ of the input vector to produce a new probability distribution. In addition, n represents the total number of elements in the input vector $\mathbf{x}$. $Temperature$ is used to adjust the soft objective function in the KD process to achieve a model-optimal solution between the output probability distribution of the teacher model and the output probability distribution of the student model. The output of this function, referred to as the soft labels,

is used to calculate the soft loss with the student model soft predictions.

$$Softmax(\mathbf{x}_i) = \frac{\exp\left(\frac{\mathbf{x}_i}{Temperature}\right)}{\sum_{j=1}^{n} \exp\left(\frac{\mathbf{x}_j}{Temperature}\right)} \tag{5.1}$$

The student model processes the input data in two ways. One process is similar to standard neural network training, where the model's output, referred to as the hard predictions, is compared with the ground truth to calculate the hard loss. The other process involves passing the model output through Equation (5.1) for feature amplification, resulting in soft predictions.

The student model produces two types of output from the input medical images: soft predictions and hard predictions. The soft predictions are obtained from the final Softmax layer of the student model and are used to calculate the Soft Loss by comparing them with the soft labels from the pre-trained teacher model. The hard predictions are compared with the ground truth labels to compute the Hard Loss.

The total loss used to update the parameters of the student model is a weighted sum of these two losses, as formulated in Equation (5.2):

$$\text{TotalLoss} = \alpha \times \text{SoftLoss} + (1 - \alpha) \times \text{HardLoss} \tag{5.2}$$

Here, α is a hyperparameter in the range $(0, 1)$, controlling the contribution of each type of loss to the total loss. The value of α is empirically determined to effectively balance Soft Loss and Hard Loss.

This implementation of KD in the FKD-Med framework allows for efficient training of models with reduced computational requirements while maintaining a high level of performance.

Table 5.1: Comparison of Parameter Quantities Between Student Model and Teacher Model

Model	**Parameter Counts**	**Parameter Optimization**
tiny-unet [4]	102498	1×
Unet++ [18]	9162786	89×
ResUNet [5]	13040770	127×
AttentionUnet [77]	34877486	340×
TransUnet [9]	105322146	1027×

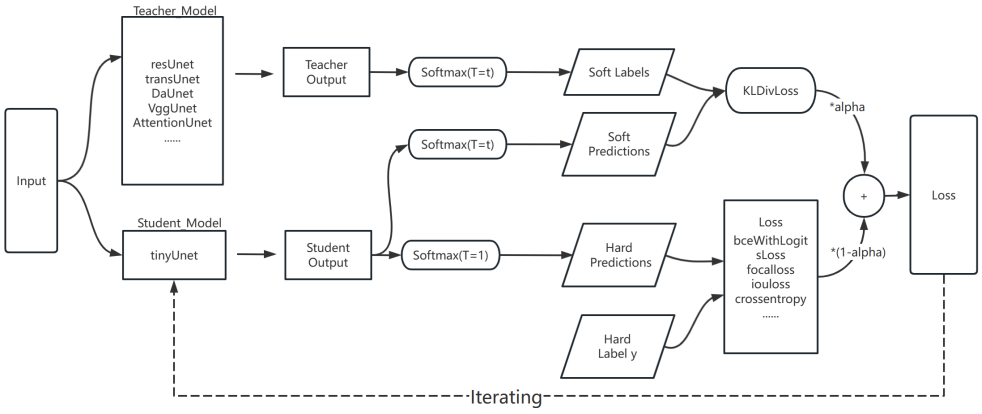

Figure 5.4: Detailed illustration of the Knowledge Distillation (KD) process in FKD-Med.

Table 5.2: Comparative of FKD-Med's Communication Efficiency, and Data Privacy with Related Models

Model	FL	KD	Parameter Optimization
Spirit Distill [91]	No	Yes	2.48×
ContextNet [92]	No	Yes	20×
MKANet [93]	No	Yes	2×
FedUKD [81]	Yes	Yes	62×
FKD-Med(ResUnet)	Yes	Yes	127×
FKD-Med(TransUnet)	Yes	Yes	1027×

5.2.4 U-Net-like Model and Loss Function in FKD-Med

The third key component of the FKD-Med framework is the U-Net Models Library. This library houses a diverse collection of U-Net-like models, each uniquely suited for medical image segmentation tasks. The models, such as U-Net [4], ResUNet [5], TransUnet [9], and others, can be selected and customized to meet the specific requirements of the task at hand. One of the distinguishing features of the framework is its dynamic nature. As research progresses and new models emerge, this study continually update the U-Net Models Library, ensuring that the users have access to the most advanced and effective tools for their segmentation tasks. The inclusion of a diverse range of U-Net-like models and customizable loss functions is a novel aspect

of FKD-Med. This flexibility is unprecedented and allows the framework to be adapted for a wide range of medical tasks beyond image segmentation.

The Table 5.1 provides a compelling comparison between the parameter volumes of distinct models, specifically between more intricate Teacher Models and their simpler Student Model counterparts. Remarkably, the parameter counts in the Teacher Models are multiplied by factors of 89, 127, 340, and 1027 when compared to the Student Models. This stark contrast underscores the efficiency of employing KD techniques when training the Tiny-Unet model. This technique plays a pivotal role in condensing the model size without significant loss in performance. Furthermore, this approach proves to be an asset in the context of FL over medical datasets, where it substantially trims communication overheads. Thus, KD emerges as a game changer, enhancing computational efficiency and enabling more effective model deployment in resource-constrained scenarios.

The Table 5.2 compares FKD-Med's communication efficiency and data privacy with other models. It details how FKD-Med applies FL and KD in its ResUnet and TransUnet variants to achieve marked parameter optimization. Compared to models such as U-net, Spirit Distill [91], ContextNet [92], MKANet [93], and FedUKD [81], FKD-Med(ResUnet) demonstrates a 127-fold increase in parameter efficiency, and FKD-Med(TransUnet) achieves an even more impressive 1027-fold enhancement. These figures highlight FKD-Med's significant strides in optimizing communication efficiency and reinforcing data privacy.

The fourth cornerstone of the FKD-Med framework is the provision of a variety of Loss Functions. This feature equips the framework with multiple loss functions that can be utilized to train the models. These loss functions, which include but are not limited to BCELoss, DiceLoss, and Tversky Loss, can be selected based on the nature of the segmentation task and the type of images being processed [94]. This flexibility allows users to tailor their choice of loss function to the specific requirements of their task, thereby optimizing the performance of their model training process.

5.3 Case Study and Experimental Analysis

In this case study, this study conducted medical image segmentation experiments on two datasets to validate the effectiveness of FKD-Med. As depicted in the Fig 5.2, this study virtualized three hospitals as nodes within the FL architecture. In the KD phase, a two-layer Tiny-Unet was employed as the student model, with both ResUNet and TransUnet serving as teacher models.

The experimental results affirm the efficacy of FKD-Med in two significant

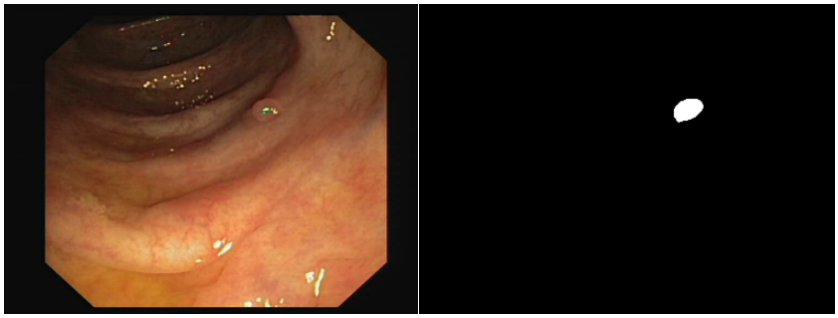

Figure 5.5: Polyp images and corresponding labels – CVC-ClinicDB Dataset

dimensions. Firstly, in terms of model lightweighting, the results demonstrate that KD under FL can substantially reduce the model size. The student models' parameters were reduced to 1/127 and 1/1027 of the teacher models', respectively. Secondly, concerning accuracy computation, the models subjected to KD exhibited accuracy improvements of 0.25%, 0.43%, 1.35%, and 1.46% respectively, given the same parameter volume. These findings not only confirm the practicality of FKD-Med but also underline its potential to enhance both efficiency and precision in the context of medical computations.

5.3.1 Datasets

Two medical datasets from two different open-source data websites is used to demonstrate the joint learning of individual Unet variant models in using KD. Specific details are given below:

This studyused two different types of medical image data to demonstrate the applicability of the FKD-Med model:

5.3.1.1 CVC-ClinicDB Dataset

The CVC-ClinicDB dataset [85] consists of 612 images of polyps and corresponding ground truth binary segmentation masks of standard resolution 384x288. CVC-ClinicDB [85] is the official database to be used in the training stages of MICCAI 2015 Sub-Challenge on Automatic Polyp Detection Challenge in Colonoscopy Videos.

To accommodate federal learning scenarios, the CVC-ClinicDB dataset was divided into the following 2 datasets. The CVC-ClinicDB dataset was divided into the following 2 datasets:

- Training - A total of 573 raw medical data images were randomized into three groups

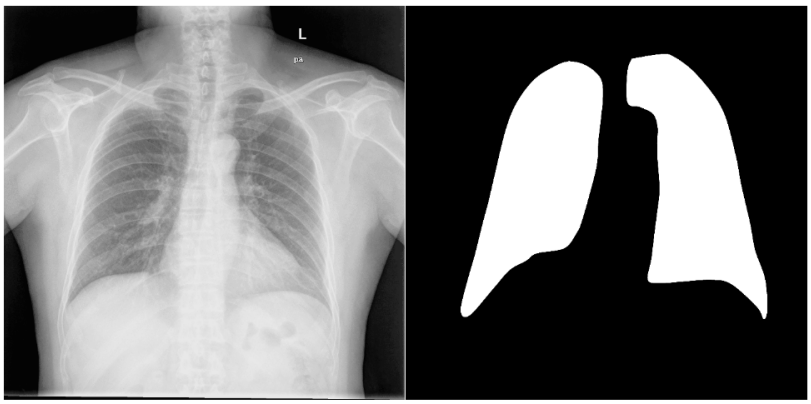

Figure 5.6: X-rays and corresponding masks – Chest-Xray Dataset

- Testing - 141 raw medical data images in total

5.3.1.2 Chest Xray Masks and Labels Dataset

The Chest-Xray dataset [86] [87] consists of 612 images of polyps and corresponding ground truth binary segmentation masks of standard resolution 384x288. To accommodate federal learning scenarios, the Chest-Xray dataset was divided into the following 2 datasets. The Chest Xray dataset was divided into the following 2 datasets:

- Training - A total of 704 raw medical data images were randomized into three groups
- Testing - 563 raw medical data images in total

5.3.2 Evaluation Metrics

In the evaluation of the proposed medical image segmentation approach, a critical metric employed is pixel-level accuracy. This metric offers a granular assessment of segmentation quality by examining the individual pixel predictions. Specifically, each pixel on the predicted segmentation map is classified as 1 or 0, representing the two distinct classes of interest in the image. The pixel-level accuracy is then calculated as the ratio of correctly classified pixels to the total number of pixels in the image [95]. Mathematically, the pixel-level accuracy (PA) can be expressed as (5.3):

$$PA = \frac{\text{Number of Correct Pixels}}{\text{Total Number of Pixels}} = \frac{\sum_{i=1}^{N} \delta(p_i, g_i)}{N} \tag{5.3}$$

where p_i is the predicted value of the i-th pixel, g_i is the ground truth value of the i-th pixel, N is the total number of pixels in the image, and $\delta(x, y)$ is the Kronecker delta function, equal to 1 if $x = y$ and 0 otherwise. This metric encapsulates the exactness of the segmentation, providing a robust and straightforward measure to evaluate the model's performance on individual medical images.

5.3.3 Experimental Setup

A FL architecture for image segmentation has been successfully implemented using the Flower framework [89], specifically designed for medical image data segmentation. To demonstrate the performance of the framework, this study used three client nodes in the experiments, with each node aimed at representing a different hospital site, simulating a diverse and realistic FL environment. However, in practical applications, the segmentation of medical image data involves the collaborative contributions of thousands of hospital models. This complexity inherently leads to high communication costs, underscoring the challenges and the necessity of efficient tools like Flower in the real-world deployment of such systems.

In the KD process, the teacher models were specifically selected as ResUNet and TransUnet, both of which are state-of-the-art (SOTA) models in medical image segmentation. ResUNet is an improved U-net architecture, boasting a total of 13,040,770 parameters, and integrates residual connections to overcome the vanishing gradient problem [5]. This design facilitates deeper network training and seamlessly combines the strengths of the U-net structure with residual networks, offering improved feature extraction and model generalization. On the other hand, TransUnet represents a fusion of transformer and U-net architectures, comprising 105,322,146 parameters [9]. Capitalizing on the flexibility and attention mechanisms of Transformers, TransUnet's unique amalgamation enables precise localization and rich contextual information, making it highly effective for various segmentation tasks.

The student model chosen in this study differs from the unconventional U-net model. This variant of the U-net is derived by taking the initial U-net structure [4] and reducing both the number of layers and the parameters within the structure. The proposed model consists of two upper and lower sample layers each, two layers fewer than the initial U-net. The filters of the convolution module have been reduced to 32 and 64, instead of the 64, 128, 256, and 512 filters found in the initial U-net. The reduction in the number of layers and the scaling down of the filters led to the creation of Tiny-Unet, with a total of 102,498 parameters.

In stand-alone training, each user will independently complete 50 training

sessions and 50 tests to correspond to the FL Framework. In FL training and in FL training for KD according to FedAVG characteristics, a complete federation is completed by 10 training sessions and one test. This process is repeated five times to complete the training of the whole model.

During the training process, three Nvidia RTX 3090 GPUs were utilized as computing devices, customized to the number of clients, and integrated with the PyTorch framework. For the training of the student model using KD, specific parameters were optimized. The distillation temperature *Temperature* was set at 5, and the proportion of training loss transferred from the teacher model to the student model was calibrated to 0.5 for the Chest X-ray Datasets and 0.8 for the CVC-ClinicDB datasets, reflecting the distinctive characteristics of each dataset.

5.3.4 Experimental Results

The overall experimental results were divided into three distinct categories: models trained without FL, models trained with FL but without KD, and models trained with both FL and KD. This division served to comprehensively assess the impact of each technique. Currently, to underscore the efficacy of KD, parallel experiments were conducted on variants of U-net, including Tiny-Unet, ResUNet, and TransUnet. These experiments adhere to the following comprehensive assessment criteria:

- Baseline Performance Metrics: The initial aspect evaluates the fundamental effectiveness of various models, including the FKD-Med framework, in different configurations: without FL, with FL but without KD, and with both FL and KD. This provides a foundational landscape for understanding the isolated and synergistic impacts of FL and KD within FKD-Med. The detailed results of these evaluations are presented in Table 5.3, Table 5.6, Fig. 5.7 and Fig. 5.9.
- Robustness and generalizability: As the second dimension, this study extend the analysis to incorporate robustness and generalizability features, particularly in the FKD-Med framework. Rigorous validation techniques such as 5-fold cross-value are employed to gauge the models' resilience and adaptability across different data splits, thus adding statistical weight to the overall results. Detailed insights from this assessment can be found in Table 5.4, Table 5.7, Fig. 5.8, and Fig. 5.10.
- Scalability and Efficiency Analysis: This final aspect emphasizes the scalability and efficiency attributes of FKD-Med, beyond mere accuracy evaluation, and meticulously scrutinizes FKD-Med's capability

to streamline models. By achieving nearly comparable accuracy rates with more lightweight architectures, FKD-Med inherently enhances communication efficiency in FL scenarios, considering variables such as training set size and parameter counts. These findings are elaborated on in Table 5.5 and Table 5.8. The experimental design, reflected in Tables 5.5 and 5.8, demonstrated the performance of FKD-Med in scenarios that prioritize data privacy and parameter efficiency, crucial for practical deployments.

These three dimensions collectively provide a complete and comprehensive understanding of the methodological strengths and potential areas for improvement.

5.3.4.1 Results on CVC-ClinicDB Datasets

Table 5.3: Comparative Evaluation of tinyUnet and FKD-Med on the CVC-ClinicDB Dataset with Identical Model Parameter Counts

Model	Teacher in KD	FL	KD	Loss	Time	Number of Parameters	Acc
tinyUnet [4]	N/A	×	×	0.2218	25min03s	102498	90.68%
tinyUnet [4] + FL	N/A	✓	×	0.2553	15min23s	102498	90.80%
tinyUnet in FKD-Med(ours)	**ResUnet** [5]	✓	✓	0.2185	19min35s	102498	**91.05%**
tinyUnet in FKD-Med(ours)	**TransUnet** [9]	✓	✓	0.2105	20min54s	102498	**91.23%**

Table 5.4: The 5-Fold Cross-Validation Accuracy Results for tinyUnet and FKD-Med Variants on the CVC-ClinicDB Dataset, Further Validating the Comparative Evaluation Under Identical Model Parameter Counts

Model	Teacher in KD	FL	KD	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
tinyUnet [4]	N/A	×	×	90.91%	91.09%	91.62%	91.50%	90.78%	91.18%
tinyUnet [4] + FL	N/A	✓	×	91.20%	91.53%	91.62%	91.50%	90.42%	91.25%
tinyUnet in FKD-Med(ours)	**ResUnet** [5]	✓	✓	91.66%	91.62%	92.23%	91.50%	91.21%	**91.64%**
tinyUnet in FKD-Med(ours)	**TransUnet** [9])	✓	✓	91.95%	91.54%	91.62%	91.51%	91.42%	**91.61%**

Table 5.5: Comparison of Parameter Counts Between Data-Scalable FKD-Med of tinyUnet Versus Non-Data-Scalable Complex Models on CVC-ClinicDB Dataset, Maintaining Similar Accuracy Levels

Model	Teacher in KD	FL	KD	Acc	Training Set Size	Parameter Counts	Parameter Optimization
ResUnet [5]	N/A	×	×	90.79%	73	13040770	-
tinyUnet in FKD-Med(ours)	**ResUnet** [5]	✓	✓	**91.05%**	489	**102498**	**1/127**
TransUnet [9]	N/A	×	×	90.79%	73	105322146	-
tinyUnet in FKD-Med(ours)	**TransUnet** [9]	✓	✓	**91.23%**	489	**102498**	**1/1027**

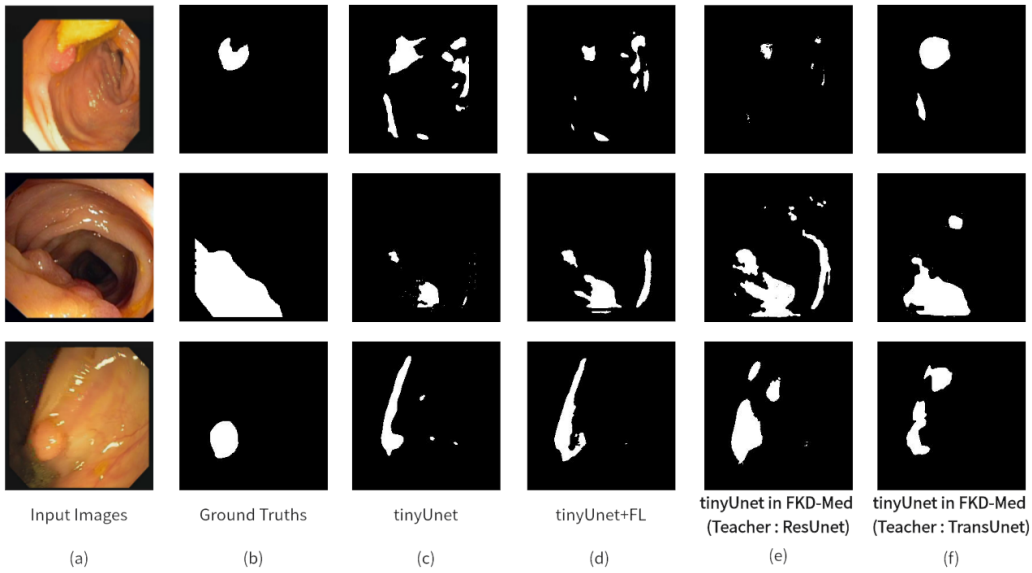

Figure 5.7: Comparative visualization of segmentation results on the CVC-ClinicDB datasets using various training models.

Table 5.3 provides a detailed comparison between Tiny-Unet and its FKD-Med variants on the CVC-ClinicDB dataset, all with identical model parameter counts. The FKD-Med models show improved performance in all metrics. Specifically, using ResUNet and TransUnet as teachers, the FKD-Med models achieve accuracies of 91.05% and 91.23%, respectively, outperforming the Tiny-Unet 90 baseline. 68%. The FKD-Med variants also register lower loss values of 0.2185 and 0.2105, compared to the baseline value of 0.2218. These results underscore the effectiveness of incorporating KD within a FL framework.

Table 5.4 presents the results of the 5-fold cross-value. The FKD-Med variants outshine their counterparts in terms of average accuracy. Specifically, the FKD-Med model trained with ResUNet and TransUnet as teachers achieved an average precision of 91. 64% and 91. 61%, respectively. These results affirm the robustness and generalizability of the model, which is particularly significant given the medical image segmentation context where high reliability is essential.

Table 5.5 highlights the superiority of the FKD-Med framework not just in terms of parameter efficiency, but also in the context of real-world applications that demand both privacy preservation and communication efficiency. FKD-Med incorporates FL to enable a decentralized training

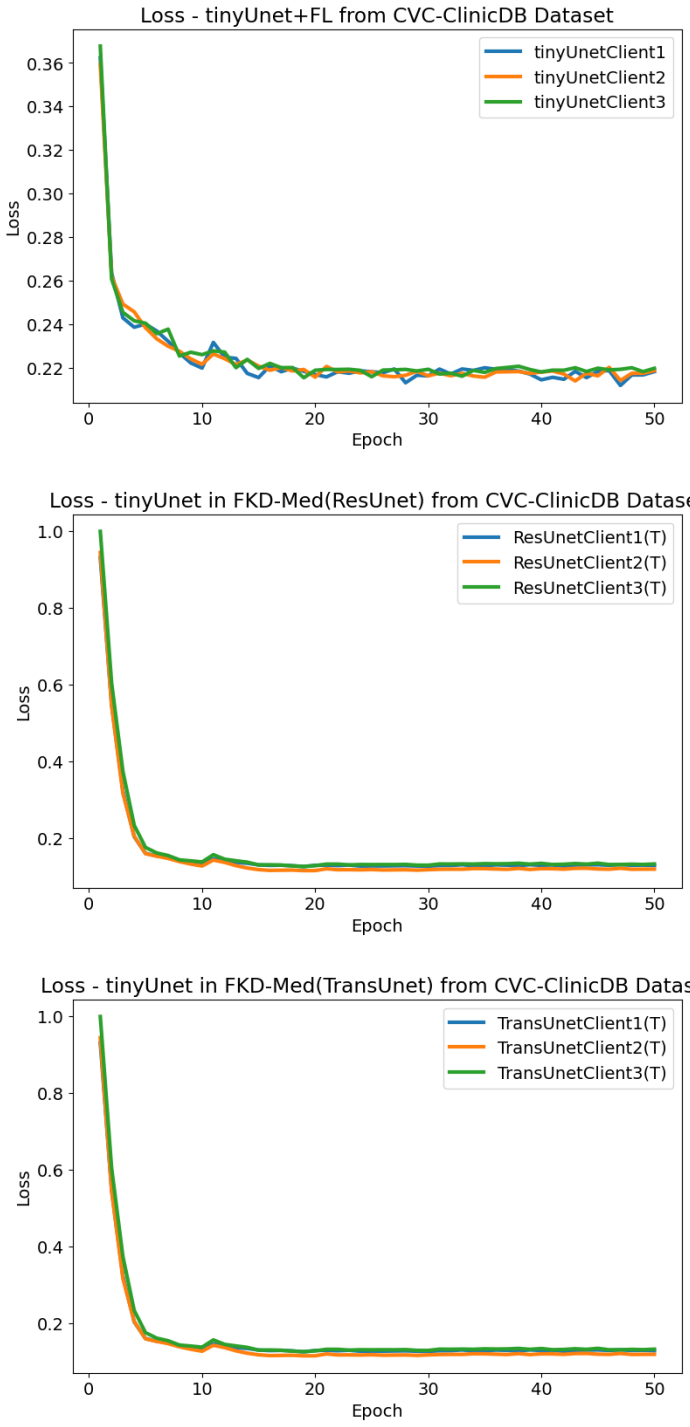

Figure 5.8: Training loss evolution on CVC-ClinicDB Datasets.

paradigm that safeguards data privacy between multiple medical institutions. This architecture naturally facilitates the collation of a larger and more diverse training dataset compared to the traditional centralized methods used in ResUnet and TransUnet. Consequently, FKD-Med benefits from a more robust learning environment. Despite the larger pool of training data, FKD-Med requires only a fraction of the model parameters - 1 / 127 and 1/1027 as compared to ResUnet and TransUnet, respectively. This drastic reduction in model size is not just a technical achievement; it has profound implications for real-world FL applications. Smaller models mean that less data need to be communicated between nodes, drastically reducing communication overhead and making it feasible to include more nodes in the network.

Fig.5.7 provides a visual representation of the experimental results in the CVC-ClinicDB dataset. This comparative visualization underscores the qualitative superiority of the FKD-Med models over their counterparts. It is evident from the segmented images that the FKD-Med variants, particularly when guided by ResUNet and TransUnet as teacher models, produce segmentations that are not only more accurate but also consistently closer to the ground truth. This visual affirmation reiterates the substantial benefits of merging FL with KD, creating an effective synergy for medical image segmentation. The depicted results offer a tangible perspective, substantiating the model's capabilities in capturing intricate morphological details, and further emphasizing the strength and robustness of the FKD-Med approach.

Fig.5.8 presents the evolution of the loss over time for three clients in the FL setting. As can be observed from the depicted trends, the FKD-Med models, specifically those that take advantage of ResUnet [5] and TransUnet [9] as teacher models, exhibit a distinct advantage in convergence speed. The loss for these models reduces more rapidly with increasing training iterations compared to the conventional tinyUnet [4] with FL. More importantly, while all models experience fluctuations in loss as training progresses, the FKD-Med variants show considerably smoother loss trajectories with less pronounced oscillations. This steadiness not only underlines the robustness of the FKD-Med framework, but also highlights its enhanced ability to resist overfitting and maintain stable learning rates across clients in the federated setup.

5.3.4.2 Results on Chest Xray Datasets

In Table 5.6, the performance of tinyUnet and the FKD-Med models on the chest X-ray dataset, each with consistent model parameter counts, is detailed. The standalone tinyUnet achieves 95.40% accuracy with a loss of 0.1321. When FL is integrated, the accuracy slightly drops to 94.24%

Table 5.6: Comparative Evaluation of tinyUnet and FKD-Med on the Chest-Xray Dataset with Identical Model Parameter Counts

Model	Teacher in KD	FL	KD	Loss	Time	Number of Parameters	Acc
tinyUnet [4]	N/A	×	×	0.1321	57min26s	102498	95.40%
tinyUnet [4] + FL	N/A	✓	×	0.1576	56min31s	102498	94.24%
tinyUnet in FKD-Med(Ours)	**ResUnet** [5]	✓	✓	0.1278	50min28s	102498	**95.59%**
tinyUnet in FKD-Med(Ours)	**TransUnet** [9]	✓	✓	0.1207	52min14s	102498	**95.70%**

with a loss of 0.1576. However, the FKD-Med models excel in this context: the version using ResUnet as the teacher gains 95. 59% precision with a 0.1278 loss, while the one paired with TransUnet boasts an accuracy of 95. 70% and a minimal loss of 0.1207. Furthermore, FKD-Med models optimize training time, resulting in 50 min28 and 52 min14 for ResUnet and TransUnet variants, respectively. This reaffirms the FKD-Med framework's prowess, particularly in FL environments on the chest-X-ray dataset.

Table 5.7: The 5-Fold Cross-Validation Accuracy Results for tinyUnet and FKD-Med Variants on the Chest-Xray Dataset, Further Validating the Comparative Evaluation Under Identical Model Parameter Counts

Model	Teacher in KD	FL	KD	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
tinyUnet [4]	N/A	×	×	94.34%	94.49%	93.28%	93.80%	93.44%	93.87%
tinyUnet [4] + FL	N/A	✓	×	95.01%	94.57%	94.45%	95.07%	94.79%	94.78%
tinyUnet in FKD-Med	**ResUnet** [5]	✓	✓	95.56%	94.99%	96.68%	95.15%	96.49%	**95.77%**
tinyUnet in FKD-Med	**TransUnet** [9]	✓	✓	96.74%	95.14%	95.22%	95.15%	96.72%	**95.80%**

In Table 5.7, this study delved deeper into the robustness and generalizability of the models, particularly the FKD-Med framework, by employing a 5-Fold Cross-Validation on the Chest-Xray dataset. The results reiterate the superior performance of FKD-Med, while the standalone tinyUnet averages 93. 87%, the FKD-Med with ResUnet and TransUnet teachers achieve averages of 95.77% and 95.80%, respectively. These consistent results across different data splits highlight the resilience and adaptability of FKD-Med, emphasizing its statistical significance in the evaluation.

Table 5.8: Comparison of Parameter Counts Between Data-Scalable FKD-Med of tinyUnet Versus Non-Data-Scalable Complex Models on Chest-Xray Dataset, Maintaining Similar Accuracy Levels

Model	Teacher in KD	FL	KD	Acc	Training Set Size	Parameter Counts	Parameter Optimization
ResUnet [5]	N/A	×	×	95.84%	96	13040770	-
tinyUnet in FKD-Med(ours)	**ResUnet** [5]	✓	✓	95.59%	563	**102498**	**1/127**
TransUnet [9]	N/A	×	×	97.38%	96	105322146	-
tinyUnet in FKD-Med(ours)	**TransUnet** [9]	✓	✓	95.70%	563	**102498**	**1/1027**

In Table 5.8, this study present a comparison between the data-scalable

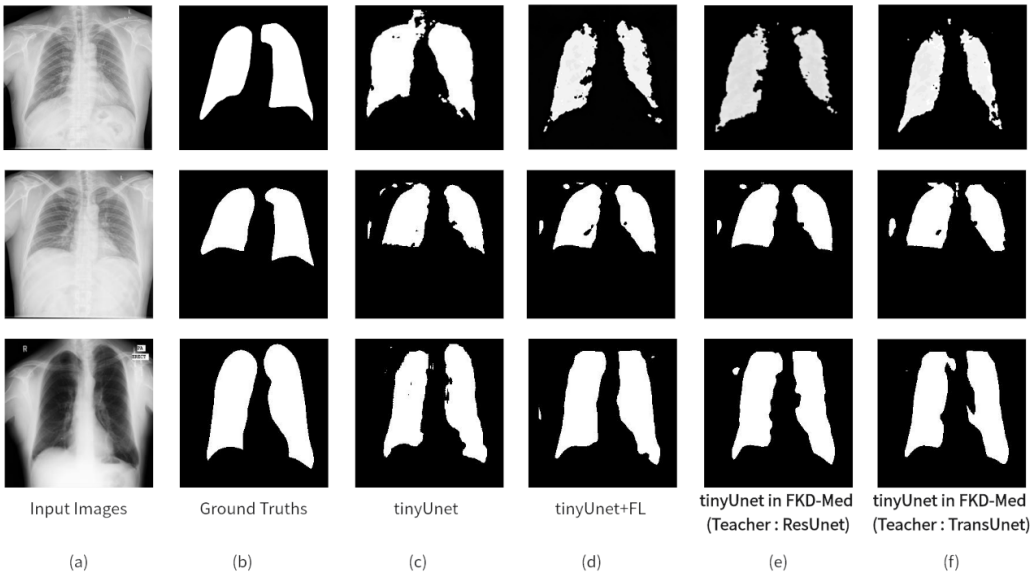

Figure 5.9: Comparative visualization of segmentation results on the the Chest Xray datasets using various training models.

FKD-Med framework applied to tinyUnet and the non-data-scalable complex models, specifically ResUnet and TransUnet, on the Chest-Xray dataset. Notably, while ResUnet and TransUnet achieve accuracies of 95. 84% and 97. 38%, respectively, with a training set size of 96, the proposed FKD-Med tinyUnet variants achieve competitive accuracies of 95.59% and 95.70% with a significantly larger training set size of 563. Most strikingly, the tinyUnet variants in the FKD-Med setting demonstrate a dramatic reduction in the number of parameters, approximately 1/127th and 1/1027th of ResUnet and TransUnet, respectively. This underlines the efficiency and scalability of the FKD-Med framework, delivering comparable performance with a fraction of the model complexity.

In Fig. 5.9, a detailed visual evaluation of the chest X-ray data set showcases the prowess of various training methodologies. Starting with the foundational X-ray images in (a) that act as the consistent input across all models, the benchmark segmentation is highlighted in (b) as Ground Truths. When observing the inherent performance of Tiny-Unet in (c), a clear distinction emerges in (d), where FL augments its capabilities. However, the standout results are evident in (e) and (f): Tiny-Unet, when synergized with teacher models ResUnet and TransUnet, respectively, under the FKD-Med framework, delivers segmentation results that underline the

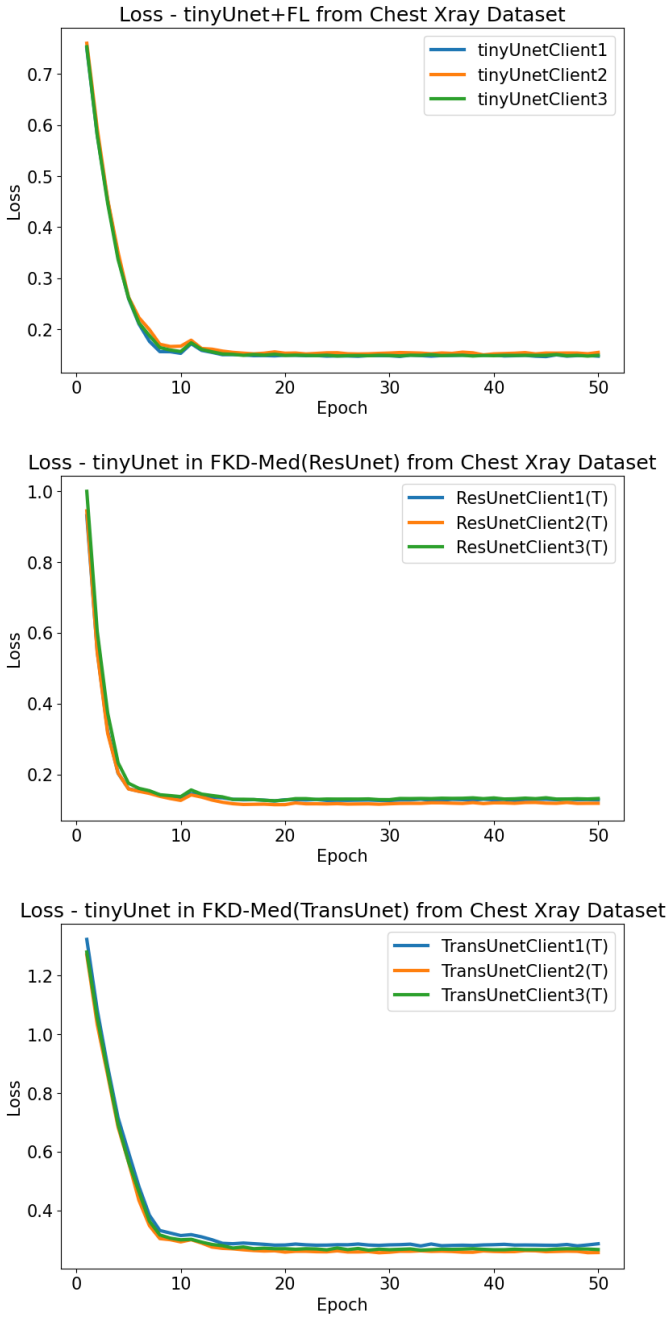

Figure 5.10: Training loss evolution on Chest-Xray Datasets.

pivotal enhancement achieved by harmoniously integrating FL with KD. This consolidation distinctly underscores FKD-Med's significant contribution to advancing segmentation accuracy and model efficiency.

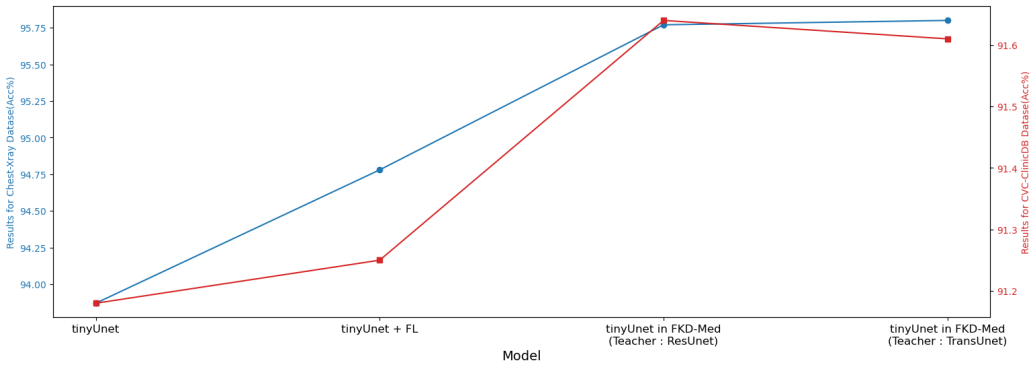

Figure 5.11: Line graph of accuracy performances for four models across CVC-ClinicDB and Chest-Xray datasets.

In Fig.5.10, the loss patterns for the chest-Xray data set echo the observations from Fig.5.8 for the CVC-ClinicDB data set. These trends, particularly within the FL environment for the chest X-ray dataset, underscore the prowess of the FKD-Med models. Using ResUnet [5] and TransUnet [9] as guiding teacher models, the FKD-Med variants manifest a pronounced edge in convergence efficiency, diverging from the standard trajectory of the baseline tinyUnet [4] under FL. Such consistent behavior accentuates FKD-Med's proficiency in curbing overfitting and ensuring uniform learning rates amongst clients, echoing the conclusions drawn from the CVC dataset.

5.3.4.3 Comprehensive Results Summary

Fig. 5.11 delineates the accuracy performances of four models in two data sets. For the Chest-Xray dataset, represented by the blue line, this study observes a notable augmentation in accuracy from the baseline tinyUnet model, peaking when adopting FKD-Med with TransUnet as the teacher. Performance in the CVC-ClinicDB dataset remains relatively stable across different model configurations, showing a modest improvement when integrated with FKD-Med, and the efficacy of FKD-Med, especially with advanced teacher models, becomes evident in elevating model performance on complex imaging tasks.

The extensive experiments demonstrate the superior performance of the FKD-Med framework over conventional methods in terms of segmentation accuracy, training efficiency, and data privacy, confirming its effectiveness and practicality in real-world medical image segmentation tasks.

5.4 Discussion

In this experiment, the joint learning of individual Unet variant models using KD was demonstrated on two medical datasets. The applicability of the FKD-Med framework was thus validated. The current challenges in medical image analysis lie in the high costs and complexity of communication. By introducing a teacher-student KD method, this study successfully optimized communication costs and reduced communication time. Existing research methods in medical image segmentation often suffer from high computational complexity, limited adaptability to diverse data, and suboptimal performance in real-world scenarios. In contrast, this experiment showcased a novel approach to handling medical image data via FL and KD. It combined FL with KD, emphasizing the utility of KD to overcome the mentioned disadvantages, thereby optimizing communication costs and improving efficiency.

5.4.1 Combination Benefits for Segmentation Challenges

In response to the challenges of data insufficiency and privacy concerns in medical image segmentation, the FKD-Med framework innovatively combines FL and KD to harness a broader spectrum of medical data without compromising patient confidentiality. This integration enables individual hospitals to contribute to a collective learning process, effectively expanding the volume of training data available across institutions. Through FL, FKD-Med aggregates insights from diverse, distributed datasets, overcoming the limitation of data scarcity at single institutions. Currently, KD compresses complex models into more efficient versions, maintaining high accuracy and significantly reducing communication load and computational demand. This approach not only addresses the challenge of limited annotated medical images, but also ensures efficient model training and deployment in a privacy-preserving manner, demonstrating a pragmatic advancement in medical image segmentation.

5.4.2 Performance Analysis of Different U-Net-like models: ResUNet vs. TransUNet in FKD-Med

TransUnet and ResUnet are different U-Net-like models tested in FKD-Med. TransUnet, as a teacher model in FKD-Med, exhibited a notable improvement in segmentation accuracy. In the experiments, TransUnet as Teacher Model achieved an average Dice coefficient of 91.23% on the CVC-ClinicDB and 95. 7% on the chest X-ray datasets, surpassing traditional Tiny

U-Net. ResUnet demonstrated enhanced training stability and efficiency. It achieved an average Dice coefficient of 91. 05% in the CVC-ClinicDB and 95. 59% in the chest X-ray datasets. In the FKD-Med framework, ResUnet achieves a parameter optimization ratio of 1/127, while TransUnet reaches an even more impressive ratio of 1/1027.

From the above analysis, this study can conclude that TransUnet not only demonstrates higher accuracy, but also achieves a greater degree of parameter optimization. This indicates that through FKD-Med, more complex models like TransUnet manifest significant advantages in both computational efficiency and accuracy. The framework's capacity to effectively distill and federate knowledge across diverse datasets enhances the performance of sophisticated architectures, making them more viable for practical applications. This synergy underscores FKD-Med's strength in leveraging complexity to yield superior segmentation results while optimizing computational resources, highlighting its potential to advance medical image analysis through the integration of advanced AI models.

5.4.3 Potential Application of FKD-Med in in real-world scenario

In real-world applications, healthcare facilities often face the challenges of data privacy, varying data volumes, and computational resource limitations. The FKD-Med architecture inherently addresses these issues by enabling collaborative learning without direct data sharing, thus preserving patient confidentiality. Furthermore, the framework's use of KD optimizes model performance by distilling knowledge from complex models into more compact, efficient representations. This process reduces the computational load on individual institutions, making advanced segmentation techniques accessible even to facilities with limited processing capabilities.

Moreover, the adaptability of FKD-Med to diverse datasets and its ability to maintain high segmentation accuracy under federated conditions demonstrate its potential for widespread adoption. For example, hospitals with smaller datasets can benefit from the collective learning process, gaining insights from larger, more diverse datasets without compromising data security. This collaborative approach not only improves the robustness of the model, but also facilitates a more inclusive healthcare research ecosystem, where institutions of varying sizes and capacities can contribute and benefit from shared advancements in medical imaging technologies.

In essence, FKD-Med stands as a beacon for the future of medical image analysis, where data privacy, computational efficiency, and collaborative

innovation converge to advance patient care. Its real-world applicability extends beyond the technical realms of machine learning, embodying the potential to revolutionize how medical data is used to improve healthcare outcomes globally.

5.4.4 Teacher Model Training Considerations in FKD-Med: Balancing Data Quantity and Communication Efficiency

In the FKD-Med framework, the training of the student model relies on the guidance of the teacher model. The performance of the teacher model plays a crucial role in determining the final performance of the student model. Ideally, the teacher model should be trained on as much data as possible to achieve strong generalizability and accurate representation of knowledge. In a federated learning environment, this implies that the teacher model should leverage data from multiple participating nodes (such as hospitals). Increasing the number of nodes not only provides more diverse training samples but also helps the teacher model learn more robust feature representations.

However, increasing the number of nodes also brings about the challenge of communication overhead. In federated learning, model parameters need to be frequently exchanged among participating nodes to enable collaborative training. The more nodes there are, the higher the communication costs will be. Therefore, there is a trade-off between data quantity and communication efficiency in the training of the teacher model. One possible solution is to selectively increase the number of nodes to achieve sufficient data diversity while keeping the communication overhead within an acceptable range.

Another promising approach is to leverage pre-trained generic segmentation models to assist in the training of the teacher model. These large-scale models, which are typically trained on extensive datasets, possess powerful feature extraction and generalization capabilities. By incorporating the knowledge from these pre-trained models into the training process of the teacher model, the performance of the teacher model can be significantly enhanced while reducing the reliance on large amounts of training data. This approach can alleviate the conflict between data quantity and communication efficiency to some extent, providing more flexibility for the application of the FKD-Med framework.

In practical applications, the training strategy of the teacher model should be determined based on the specific medical image segmentation task and available computational resources. By appropriately selecting the number of

participating nodes and leveraging pre-trained generic segmentation models, the FKD-Med framework can achieve high-accuracy medical image segmentation while ensuring data privacy and communication efficiency. This flexible training strategy further highlights the value and potential of the FKD-Med framework in real-world medical applications.

5.4.5 Limitations of FKD-Med

The framework presented in this study has two main limitations. First, while the KD part of the framework can be directly applied without significant adjustments in the study of medical image segmentation, it may require fine-tuning when used for other types of medical computation. Depending on the specific model utilized, modifications to the KD operations might be necessary. Second, the design of the KD part is constrained by the placement of soft labels, which are currently set in the last layer of the model. Future improvements could facilitate computing soft labels at any layer through simple parameter settings, thereby enhancing the framework's flexibility and applicability.

In summary, this case study experimentally demonstrated the effectiveness and feasibility of the FKD-Med framework in medical image segmentation. Through the integration of KD and FL, an innovative solution was provided for communication costs and efficiency, overcoming the shortcomings found in existing research. The FKD-Med framework not only demonstrates computational efficiency, but also has significant clinical value. By allowing data integration across different healthcare settings without compromising data privacy, FKD-Med may revolutionize collaborative medical research and lead to more personalized and effective treatments.

5.5 Chapter Summary

In this chapter, this study presented FKD-Med, an innovative open-source framework that integrates Federated Learning (FL) and Knowledge Distillation (KD) to address the challenges of data privacy, communication efficiency, and model performance in medical image segmentation. The framework is designed to be adaptable to a wide range of medical applications, extending beyond image segmentation to include diagnostic analytics, treatment planning, and drug discovery.

The key contributions of FKD-Med include its pioneering application of FL and KD for medical image segmentation, which effectively reduces communication costs in deep model training while preserving data privacy.

The framework significantly lowers computation and communication costs by compressing model parameters by factors of 127 and 1027 without sacrificing accuracy, as validated through experiments on the CVC-ClinicDB and Chest Xray datasets.

The extensive experimental results demonstrate the superior performance of FKD-Med compared to conventional methods in terms of segmentation accuracy, training efficiency, and data privacy. The framework showcases its effectiveness and practicality in real-world medical image segmentation tasks, with the potential for widespread adoption in healthcare facilities facing challenges of data privacy, varying data volumes, and computational resource limitations.

The discussion section highlights the benefits of combining FL and KD in addressing segmentation challenges, such as data insufficiency and privacy concerns. The analysis of different U-Net-like models, specifically ResUNet and TransUNet, within the FKD-Med framework reveals the advantages of complex models in achieving higher accuracy and greater parameter optimization. The potential real-world applications of FKD-Med are also discussed, emphasizing its ability to facilitate collaborative learning while preserving patient confidentiality and optimizing computational resources.

Despite its significant contributions, FKD-Med has some limitations. The KD component may require fine-tuning when applied to other types of medical computation, and the current design of the KD part is constrained by the placement of soft labels. Future improvements could focus on enhancing the framework's flexibility and applicability by allowing soft labels to be computed at any layer through simple parameter settings.

In conclusion, FKD-Med represents a groundbreaking framework that leverages the synergy between FL and KD to advance medical image segmentation. Its ability to optimize communication costs, improve efficiency, and maintain high accuracy while preserving data privacy makes it a valuable tool for collaborative medical research and personalized healthcare. The framework's adaptability and potential for real-world applications underscore its significance in revolutionizing medical image analysis and advancing patient care on a global scale.

Chapter 6

Conclusion and Future Directions

6.1 Summary of Key Findings

This dissertation has focused on improving medical image segmentation by addressing three key aspects: model accuracy, data privacy, and computational efficiency. In this dissertation, novel deep learning architectures and techniques have been proposed that leverage the power of attention mechanisms, transformer models, federated learning, and knowledge distillation to tackle the challenges in medical image segmentation.

In Chapter 3, this dissertation introduced DA-TransUNet, a dual attention transformer U-Net architecture that integrates spatial and channel attention mechanisms with transformer models. Through extensive experiments on multiple benchmark datasets, this dissertation demonstrated that DA-TransUNet effectively captures fine-grained details and long-range dependencies in medical images, leading to improved segmentation accuracy compared to state-of-the-art methods. The integration of attention mechanisms and transformer models into a U-Net-like architecture has proven to be a promising approach to improve the performance of medical image segmentation models.

Chapter 4 presented MIPC-Net, a mutual inclusion mechanism for precise boundary segmentation. MIPC-Net uses complementary information from position and channel features to enhance the delineation of complex anatomical structures and small lesions. The experimental results showcased the superiority of MIPC-Net in achieving accurate boundary segmentation compared to existing methods. The mutual inclusion of position and channel information has proven to be an effective strategy to improve the precision of segmentation models, particularly in challenging scenarios with intricate boundaries.

In Chapter 5, this dissertation introduced FKD-Med, a privacy-aware and communication-optimized framework for medical image segmentation.

FKD-Med integrates federated learning and knowledge distillation techniques to enable collaborative model training between multiple institutions while preserving data privacy. The framework also improves model efficiency by distilling knowledge from complex models to lighter ones, reducing computational requirements without compromising segmentation performance. The experimental results demonstrated the effectiveness of FKD-Med in achieving accurate segmentation while ensuring data privacy and communication efficiency. The combination of federated learning and knowledge distillation has proven to be a promising approach to enabling collaborative learning to preserve privacy and optimizing the model in medical image segmentation.

6.2 Contributions to the Field

The contributions of this dissertation to the field of medical image segmentation are significant and multifaceted. Firstly, the proposed DA-TransUNet architecture advances the state-of-the-art in medical image segmentation by leveraging the power of attention mechanisms and transformer models. The integration of spatial and channel attention with transformer models in a U-Net-like architecture provides a novel and effective approach to capture fine-grained details and long-range dependencies, leading to improved segmentation accuracy. This contribution paves the way for further exploration and adoption of attention mechanisms and transformer models in medical image segmentation tasks.

Secondly, the introduction of MIPC-Net and its mutual inclusion mechanism for precise boundary segmentation address a critical challenge in medical image segmentation. By effectively combining position and channel information, MIPC-Net enhances the delineation of complex anatomical structures and small lesions, resulting in more accurate boundary segmentation. This contribution has the potential to improve the precision and reliability of segmentation models in clinical applications, aiding in treatment planning and surgical interventions.

Third, the proposed FKD-Med framework addresses the important issues of data privacy and computational efficiency in medical image segmentation. By integrating federated learning and knowledge distillation techniques, FKD-Med enables collaborative learning preserving privacy between multiple institutions, overcoming the barriers posed by data sharing restrictions and privacy concerns. Moreover, the framework optimizes model efficiency through knowledge distillation, reducing computational requirements without sacrificing segmentation performance. This contribution has significant implications for the practical deployment of medical image segmentation

models in resource-constrained clinical settings and for facilitating multi-institutional collaborations in medical research.

In general, the contributions of this dissertation advance the field of medical image segmentation by proposing novel architectures, mechanisms, and frameworks that address key challenges related to model accuracy, data privacy, and computational efficiency. These contributions have the potential to improve patient care by allowing more accurate and efficient disease quantification, prognosis assessment, and treatment evaluation.

6.3 Recommendations for Future Research

While this dissertation has made significant contributions to the field of medical image segmentation, there are still several avenues for future research that can further enhance the performance, privacy, and efficiency of segmentation models. Some recommendations for future research include the following:

- Exploring the integration of additional attention mechanisms and transformer variants into the proposed DA-TransUNet architecture to further improve the capture of fine-grained details and long-range dependencies. Investigating the effectiveness of different attention mechanisms and transformer configurations could lead to even higher segmentation accuracy and robustness.
- Extending the MIPC-Net mechanism to handle 3D medical images and volumetric segmentation tasks. Adapting the mutual inclusion of position and channel information to the 3D domain could potentially improve the accuracy of boundary segmentation in complex anatomical structures and enable more precise quantification of lesions and organs.
- Investigating the integration of differential privacy techniques into the FKD-Med framework to provide stronger privacy guarantees and enhance the protection of sensitive patient information. Exploring the trade-offs between privacy, model performance, and communication efficiency in the context of federated learning and knowledge distillation could lead to more secure and practical solutions for collaborative learning in medical image segmentation.
- Developing advanced model compression and acceleration techniques to further improve the computational efficiency of medical image segmentation models. Investigating the use of pruning, quantization, and other optimization techniques in conjunction with knowledge distillation could enable the deployment of highly accurate segmentation models on resource-constrained devices and real-time clinical applications.

- Exploring the generalizability and transferability of the proposed methods and frameworks to different medical imaging modalities and anatomical regions. Evaluating the effectiveness of DA-TransUNet, MIPC-Net, and FKD-Med on a wider range of medical image segmentation tasks, such as brain tumor segmentation, cardiac segmentation, and retinal vessel segmentation, could demonstrate the broad applicability and robustness of these approaches.
- Investigating the interpretability and explainability of the proposed segmentation models. Developing techniques to visualize and understand the decision-making process of attention mechanisms and transformer models could enhance the trust and adoption of these methods in clinical practice, facilitating the collaboration between medical experts and AI systems.

By addressing these future research directions, this dissertation can further advance the field of medical image segmentation, bringing us closer to the goal of accurate, privacy-preserving, and efficient segmentation models that can be seamlessly integrated into clinical workflows. The continued development and refinement of these techniques has the potential to revolutionize medical image analysis and ultimately improve patient care by enabling a more precise diagnosis, treatment planning, and monitoring of various diseases and conditions.

References

[1] G. Sun, H. Shu, F. Shao, T. Racharak, W. Kong, Y. Pan, J. Dong, S. Wang, L.-M. Nguyen, and J. Xin, "Fkd-med: Privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation," *IEEE Access*, 2024.

[2] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. 1398237, 2024.

[3] G. Li, D. Jin, Q. Yu, and M. Qi, "Ib-transunet: Combining information bottleneck and transformer for medical image segmentation," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 3, pp. 249–258, 2023.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[6] 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.

[7] Z. Shi, C. Miao, U. J. Schoepf, R. H. Savage, D. M. Dargis, C. Pan, X. Chai, X. L. Li, S. Xia, X. Zhang *et al.*, "A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images," *Nature communications*, vol. 11, no. 1, p. 6090, 2020.

[8] D. Maji, P. Sigedar, and M. Singh, "Attention res-unet with guided decoder for semantic segmentation of brain tumors," *Biomedical Signal Processing and Control*, vol. 71, p. 103077, 2022.

[9] 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.

[10] A. Jamali, S. K. Roy, J. Li, and P. Ghamisi, "Transu-net++: Rethinking attention gated transu-net for deforestation mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103332, 2023.

[11] 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*. Springer, 2022, pp. 205–218.

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

[13] 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*, vol. 71, pp. 1–15, 2022.

[14] Y. Yang and S. Mehrkanoon, "Aa-transunet: Attention augmented transunet for nowcasting tasks," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 01–08.

[15] 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*. Springer, 2016, pp. 179–187.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[18] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[19] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[20] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 605132, 2020.

[21] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional convlstm u-net with densley connected convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International conference on machine learning*. PMLR, 2015, pp. 1462–1471.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[25] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 299–307.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[28] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1236–1242.

[29] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[30] Y. Cai, H. Li, J. Xin, and G. Sun, "Mlda-unet: Multi level dual attention unet for polyp segmentation," in *2022 16th ICME International Conference on Complex Medical Engineering (CME)*. IEEE, 2022, pp. 372–376.

[31] 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*, vol. 30, 2017.

[32] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.

[33] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns 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*. Springer, 2021, pp. 14–24.

[34] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: Residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[35] B. C. Tedeschini, S. Savazzi, R. Stoklasa, L. Barbieri, I. Stathopoulos, M. Nicoli, and L. Serio, "Decentralized federated learning for healthcare

networks: A case study on tumor segmentation," *IEEE Access*, vol. 10, pp. 8693–8708, 2022.

[36] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, p. 12598, 2020.

[37] F. Ullah, M. Nadeem, M. Abrar, F. Amin, A. Salam, and S. Khan, "Enhancing brain tumor segmentation accuracy through scalable federated learning with advanced data privacy and security measures," *Mathematics*, vol. 11, no. 19, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/19/4189

[38] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "Fedmix: Mixed supervised federated learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[39] J. Miao, Z. Yang, L. Fan, and Y. Yang, "Fedseg: Class-heterogeneous federated learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8042–8052.

[40] 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*, vol. 34, 2020, pp. 775–783.

[41] Z. Tian, P. Chen, X. Lai, L. Jiang, S. Liu, H. Zhao, B. Yu, M.-C. Yang, and J. Jia, "Adaptive perspective distillation for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1372–1387, 2022.

[42] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[43] D. Ji, H. Wang, M. Tao, J. Huang, X.-S. Hua, and H. Lu, "Structural and statistical texture knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 876–16 885.

[44] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 319–12 328.

[45] X. Li, B. Chen, and W. Lu, "Feddkd: Federated learning with decentralized knowledge distillation," *Applied Intelligence*, pp. 1–17, 2023.

[46] Z. Wu, S. Sun, Y. Wang, M. Liu, Q. Pan, X. Jiang, and B. Gao, "Fedict: Federated multi-task distillation for multi-access edge computing," *IEEE Transactions on Parallel and Distributed Systems*, 2023.

[47] Y. Chen, W. Lu, X. Qin, J. Wang, and X. Xie, "Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[48] T.-O. Tran and N. Q. K. Le, "Sa-ttca: An svm-based approach for tumor t-cell antigen classification using features extracted from biological sequencing and natural language processing," *Computers in Biology and Medicine*, p. 108408, 2024.

[49] N. Q. K. Le, "Hematoma expansion prediction: still navigating the intersection of deep learning and radiomics," *European Radiology*, pp. 1–3, 2024.

[50] T.-O. Tran, T. H. Vo, and N. Q. K. Le, "Omics-based deep learning approaches for lung cancer decision-making and therapeutics development," *Briefings in Functional Genomics*, p. elad031, 2023.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[52] 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*, vol. 30, 2017.

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

[54] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[55] P. Tang, C. Zu, M. Hong, R. Yan, X. Peng, J. Xiao, X. Wu, J. Zhou, L. Zhou, and Y. Wang, "Da-dsunet: dual attention-based dense su-net for automatic head-and-neck tumor segmentation in mri images," *Neurocomputing*, vol. 435, pp. 103–113, 2021.

[56] S. Sahayam, R. Nenavath, U. Jayaraman, and S. Prakash, "Brain tumor segmentation using a hybrid multi resolution u-net with residual dual attention and deep supervision on mr images," *Biomedical Signal Processing and Control*, vol. 78, p. 103939, 2022.

[57] B. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. Langerak, and A. Klein, "Segmentation outside the cranial vault challenge," in *MICCAI: Multi Atlas Labeling Beyond Cranial Vault-Workshop Challenge*, 2015.

[58] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[59] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[60] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[61] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*. Springer, 2020, pp. 451–462.

[62] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *MultiMedia Modeling:*

27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. Springer, 2021, pp. 218–229.

[63] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 233–245, 2013.

[64] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[65] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 2441–2449.

[66] R. Azad, M. T. Al-Antary, M. Heidari, and D. Merhof, "Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model," *IEEE Access*, vol. 10, pp. 108 205–108 215, 2022.

[67] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[69] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[70] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[71] A. Gupta, R. Gupta, S. Gehlot, and S. Goswami, "Segpc-2021: Segmentation of multiple myeloma plasma cells in microscopic images," *IEEE Dataport*, vol. 1, no. 1, p. 1, 2021.

[72] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[73] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.

[74] J. S. Suri, M. Bhagawati, S. Agarwal, S. Paul, A. Pandey, S. K. Gupta, L. Saba, K. I. Paraskevas, N. N. Khanna, J. R. Laird *et al.*, "Unet deep learning architecture for segmentation of vascular and non-vascular images: A microscopic look at unet components buffered with pruning, explainable artificial intelligence, and bias," *Ieee Access*, vol. 11, pp. 595–645, 2022.

[75] S.-T. Tran, M.-H. Nguyen, H.-P. Dang, and T.-T. Nguyen, "Automatic polyp segmentation using modified recurrent residual unet network," *IEEE Access*, vol. 10, pp. 65 951–65 961, 2022.

[76] Z.-J. Gao, Y. He, and Y. Li, "A novel lightweight swin-unet network for semantic segmentation of covid-19 lesion in ct images," *Ieee Access*, vol. 11, pp. 950–962, 2022.

[77] S. Lian, Z. Luo, Z. Zhong, X. Lin, S. Su, and S. Li, "Attention guided u-net for accurate iris segmentation," *Journal of Visual Communication and Image Representation*, vol. 56, pp. 296–304, 2018.

[78] Y. Cai, H. Li, J. Xin, and G. Sun, "Mlda-unet: Multi level dual attention unet for polyp segmentation," in *2022 16th ICME International Conference on Complex Medical Engineering (CME)*. IEEE, 2022, pp. 372–376.

[79] M.-H. Sheu, S. S. Morsalin, S.-H. Wang, L.-K. Wei, S.-C. Hsia, and C.-Y. Chang, "Fhi-unet: faster heterogeneous images semantic segmentation design and edge ai implementation for visible and thermal images processing," *IEEE Access*, vol. 10, pp. 18 596–18 607, 2022.

[80] S. Han, S. Park, F. Wu, S. Kim, C. Wu, X. Xie, and M. Cha, "Fedx: Unsupervised federated learning with cross knowledge distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 691–707.

[81] R. Kanagavelu, K. Dua, P. Garai, N. Thomas, S. Elias, S. Elias, Q. Wei, L. Yong, and G. S. M. Rick, "Fedukd: Federated unet model with knowledge distillation for land use classification from satellite and street views," *Electronics*, vol. 12, no. 4, p. 896, 2023.

[82] R. W. Anwar, M. Abrar, and F. Ullah, "Transfer learning in brain tumor classification: Challenges, opportunities, and future prospects," in *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*, 2023, pp. 24–29.

[83] F. Ullah, M. Nadeem, M. Abrar, F. Amin, A. Salam, A. Alabrah, and H. AlSalman, "Evolutionary model for brain cancer-grading and classification," *IEEE Access*, vol. 11, pp. 126 182–126 194, 2023.

[84] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature communications*, vol. 13, no. 1, p. 2032, 2022.

[85] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[86] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 233–245, 2013.

[87] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[88] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[89] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[90] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[91] Z. Wu, Y. Jiang, M. Zhao, C. Cui, Z. Yang, X. Xue, and H. Qi, "Spirit distillation: A model compression method with multi-domain knowledge transfer," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2021, pp. 553–565.

[92] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach, "Contextnet: Exploring context and detail for semantic segmentation in real-time," *arXiv preprint arXiv:1805.04554*, 2018.

[93] Z. Zhang, W. Lu, J. Cao, and G. Xie, "Mkanet: An efficient network with sobel boundary loss for land-cover classification of satellite remote sensing imagery," *Remote Sensing*, vol. 14, no. 18, p. 4514, 2022.

[94] J. Xin and G. Sun, "Learn from each other: Comparison and fusion for medical segmentation loss," in *2021 7th International Conference on Computer and Communications (ICCC)*. IEEE, 2021, pp. 662–666.

[95] R. Fan, Z. Wang, and Q. Zhu, "Egfnet: Efficient guided feature fusion network for skin cancer lesion segmentation," in *2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI)*, 2022, pp. 95–99.

Publications

Journal Papers Related to Dissertation

[1] **Guanqun Sun***, Han Shu, Feihe Shao, Teeradaj Racharak, Weikun Kong, Yizhi Pan, Jingjing Dong, Shuang Wang, Le-Minh Nguyen*, Junyi Xin*, "FKD-Med: Privacy-Aware, Communication-Optimized Medical Image Segmentation via Federated Learning and Model Lightweighting Through Knowledge Distillation," *IEEE Access*, vol.12, pp.33687-33704, 2024. DOI: 10.1109/ACCESS.2024.3372394

[2] **Guanqun Sun***, Yizhi Pan, Weikun Kong, Zichang Xu, Jianhua Ma, Teeradaj Racharak, Le-Minh Nguyen*, Junyi Xin* "DA-TransUNet: Integrating Spatial and Channel Dual Attention with Transformer U-Net for Medical Image Segmentation," *Frontiers in Bioengineering and Biotechnology*, Volume 12-2024. DOI: 10.3389/fbioe.2024.1398237

[3] Yizhi Pan, Junyi Xin, Tianhua Yang, Teeradaj Racharak, Le-Minh Nguyen*, **Guanqun Sun***, "A Mutual Inclusion Mechanism for Precise Boundary Segmentation in Medical Images," *arXiv preprint arXiv:2404.08201* 2024.

Additional Scholarly Works

[1] Xiaofeng Zhu, Yi Zhang, Haoru Ying, Huanning Chi, **Guanqun Sun***, Lingxia Zeng*, "Modeling epidemic dynamics using Graph Attention based Spatial Temporal networks," *PLOS ONE*, vol.19, no.7, pp.e0307159, 2024. DOI: 10.1371/journal.pone.0307159

[2] Kong Wei Kun, Xin Liu, Teeradaj Racharak, **Guanqun Sun**, Jianan Chen, Qiang Ma, Le-Minh Nguyen, "WeExt: A Framework of Extending Deterministic Knowledge Graph Embedding Models for Embedding Weighted Knowledge Graphs," *IEEE Access*, vol.11, pp.48901-48911, 2023. DOI: 10.1109/ACCESS.2023.3276319

[3] Zichang Xu, Hendra S Ismanto, Dianita S Saputri, Soichiro Haruna, **Guanqun Sun**, Jan Wilamowski, Shunsuke Teraguchi, Ayan Sengupta, Songling Li, Daron M Standley, "Robust detection of infectious

disease, autoimmunity, and cancer from the paratope networks of adaptive immune receptors," *Briefings in Bioinformatics*, vol.25, no.5, pp.bbae431, 2024. DOI: 10.1093/bib/bbae431