Advancing Medical Image Segmentation & Generalization by Capturing Global Context & Mitigating Negative Knowledge Transfer Across Multi-Source Data

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (Ph.D.)

of the

Department of Computer Science,
Brunel University London

by

Ndipenoch Nchongmaje

Principal Supervisor:	Prof Yongmin Li
Second Supervisor:	Dr Alina Miron
Submission Date:	2024-12-16

Declaration

I hereby declare that this thesis is solely completed by the candidate, Ndipenoch Nchongmaje. The original research work has not been presented for the award of any other degree in the past. Some work in it has been published previously and that is stated in the text where relevant. All sources of material have been properly acknowledged and references have been provided.

Abstract

Deep learning methods have shown significant success in detecting and segmenting diseases or pathogens in medical images. However, most of these models are trained and tested on data from the same source, resulting in poor generalizability when applied to unseen data, as often encountered in real-world scenarios. This challenge is primarily due to the domain shift problem, which occurs when there is a discrepancy in data distributions between the source (training) domain and the target (testing) domain. This shift often occurs because medical images are collected from diverse sources, modalities, and vendor machines, with varying scanning protocols and expertise levels among radiologists and annotators. Furthermore, deep learning models typically require large, annotated datasets for training. Given that annotating medical images is labor-intensive and time-consuming, the size of available datasets is often limited. While numerous small, annotated datasets exist across various medical domains, directly combining them can introduce another issue known as Negative Knowledge Transfer (NKT), where knowledge from one domain negatively impacts performance in another, particularly in multi-domain training. This research aims to address these challenges by proposing the integration of Atrous Spatial Pyramid Pooling (ASPP) and Squeeze-and-Excitation (SE) blocks to capture global contextual information in the case of specific designed architectures, and knowledge transfer and domain adapters to mitigate negative knowledge transfer in the case of diverse, multi-source data. These enhancements improve the model's segmentation and generalization performance. Three key contributions are presented:
1)**Enhancing Retinal Disease Detection, Segmentation, and Generalization with an ASPP Block and Residual Connections Across Diverse Data Sources**: We propose a novel algorithm nnUNet_RASPP, an enhanced variant of nnU-Net that incorporates an Atrous Spatial Pyramid Pooling (ASPP) block immediately after the input layer to capture global contextual information, as well as residual connections to mitigate the vanishing gradient problem, thereby improving the model's generalizability across data from diverse sources (collected using three different manufacturer devices). Additionally, we conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed. Experimented on the RETOUCH Grande Challenge dataset, and evaluation results on the hidden test set show that nnUNet_RASPP outperformed the baseline nnU-Net and state-of-the-art models by a clear margin. Also, nnUNet_RASPP is the current winner of both the online and offline phases of the competition. Additionally, nnUNet_RASPP demonstrated strong generalization on unseen datasets.
2) **Dynamic Network for Global Context-Aware Disease Segmentation in Retinal Images Using Multiple ASPP and SE Blocks**: We further explore the potential of using multiple ASPP blocks at various locations, along with Squeeze-and-Excitation (SE) blocks, within a dynamic convolutional neural network (CNN) architecture that can automatically adjust the kernel size and depth of the network based on input size. We propose a novel algorithm, Deep_ResUNet++, a dynamic CNN model that incorporates multiple ASPP and SE blocks to capture global contextual information for disease segmentation in 2D B-Scans. The use of multiple ASPP and SE blocks offer a more detailed and effective method for feature

extraction, context aggregation, and feature recalibration. Deep_ResUNet++ was evaluated on two public datasets, the AROI and Duke DME datasets, outperforming state-of-the-art algorithms by a clear margin.

3) **Enhancing Medical Image Segmentation Through Knowledge Transfer with Domain-Specific Adapters Across Diverse Data Sources**: To further enhance model generalizability, we aim to leverage the synergistic potential of multiple datasets to create a single, diverse model trained on data from various sources, covering multiple modalities, organs, and disease types, collected with different device vendors and protocols. To mitigate negative knowledge transfer, we incorporate domain knowledge adapters into the network architecture. We propose two novel algorithms: (i) MMIS-Net (MultiModal Medical Image Segmentation Network), which addresses label inconsistencies through a one-hot label space and employs a similarity fusion block for multi-source medical image segmentation. And (ii) CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-Specific Batch Normalization), which integrates Vision Transformers and CNNs with domain-specific batch normalization to improve generalization. Both algorithms were evaluated on two dataset groups. The first group, comprising 10 benchmark datasets from the Medical Segmentation Decathlon (MSD) and the RETOUCH, challenge benchmark and the second group, is the HECKTOR challenge benchmark dataset. Experimental results on the hidden test sets show that both algorithms outperformed state-of-the-art algorithms and large foundation models for medical image segmentation by a clear margin, demonstrating superior generalization on new, unseen data.

In summary, this research introduces techniques to enhance model segmentation performance and generalizability by integrating Atrous Spatial Pyramid Pooling (ASPP) and Squeeze-and-Excitation (SE) blocks for capturing global contextual information in specific designed models and domain-adaptive adapters to mitigate negative knowledge transfer on diverse, multi-source data. These methods not only improve model generalization on new, unseen data but also set new benchmarks in medical image segmentation, providing robust and generalizable solutions for real-world clinical applications.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Yongmin Li and Dr. Alina Miron, for their invaluable support, expertise, constructive feedback, and guidance throughout my research journey. Their mentorship has been instrumental in helping me grow as a researcher and refine my skills. They were always there to answer my questions, steer me in the right direction when needed, and pushing me to strive for excellence. I am especially grateful to Prof. Yongmin Li for his exceptional dedication and commitment to guiding me throughout my research. His remarkable ability to consistently provide prompt, detailed, and constructive feedback is truly commendable.

I would also like to extend my thanks to the Department of Computer Science at Brunel University London for their support and assistance throughout my PhD journey.

Additionally, I am grateful to Prof. Kezhi Wang and the EU COVER project for organizing research collaborations for me with partner universities abroad. A special thank you goes to Dr. Zhan Shu (University of Alberta, Canada) for supporting me during my time in Canada as part of this collaboration.

This accomplishment would not have been possible without the unwavering support, unconditional love, and encouragement of my family. I am profoundly thankful to my mother: Jesimia Victorine, my brother: Agboryong N. Ndipenock, my sisters : Ojongakenteng Ndipenoch and Lebsia E. Ndipenoch, my partner: Sally Dowding, and my adorable son, James Ndipenoch. I must also mention my lovable nieces and nephews. I am also thankful to my brothers-in-law, Frank Kameni and Patrick Kameni. You have all stood by me from the very beginning, showing patience and understanding throughout this challenging journey.

I would also like to acknowledge my friends and peers for the stimulating ideas we shared, the projects we collaborated on, and the memorable social gatherings and events we attended together.

Finally, I am deeply thankful to Apple as a company and to my colleagues at Apple Store White City London, especially the leadership team. Your encouragement, and support, particularly during times when I needed to travel abroad for research collaborations and conferences, were invaluable and greatly appreciated.

Contents

List of Figures

List of Tables

Abbreviations

This section provides a list of abbreviations used throughout this work, along with their corresponding meanings. The abbreviations are presented in the order of their first appearance in the text.

Abbreviations	Meaning
NFL	Nerve fiber layer
OPL	Outer plexiform layer
ONL	Outer nuclear layer
ELM	External limiting membrane
RPE	Retinal pigmented epithelium
DME	Diabetic macular edema
AMD	Age-related macular degeneration
DR	Diabetic retinopathy
RD	Retinal detachment
VEGF	vascular endothelial growth factor
OCT	Optical coherence tomography
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
UV	ultraviolet
BCC	Basal cell carcinoma
SCC	squamous cell carcinoma
RCM	Reflectance Confocal Microscopy
HFUS	High-Frequency Ultrasound
MICCAI	Medical Image Computing and Computer Assisted Intervention
CNN	Convolutional neural networks
RETOUCH	Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge
HECKTOR	HEad and neCK TumOR
DS	Dice Score
IoU	Intersection over Union
AVD	Absolute Volume Difference
nnU-Net	no-new-Net
CAD	Computer-aided detection and diagnosis
SSA	Shared-specific adapter
DB	Dual-branch
AFP	Atrous feature pyramid
STSC	Spatiotemporal Separate Convolution
MSD	Medical Segmentation Decathlon
MLP	Multi-layer perceptron
BTCV	Beyond The Cranial Vault
DCAC	Domain and Content Adaptive Convolution
GPU	Graphics Processing Unit
ROI	Regions of interest
CoTr	Convolutional Neural Network and Transformer
GAP	Global Average Pooling
AH-Net	Anisotropic Hybrid Network
SGD	Stochastic gradient descent

MedLAM	Localize Anything Model for 3D Medical Images
UAM	Unified Anatomical Mapping
MSS	Multi-Scale Similarity
DAC	Domain Adaptive Convolution
CAC	Content Adaptive Convolution
DANN	Domain Adversarial Neural Network
DoFE	Domain-Oriented Feature Embedding
BCE	Binary Cross-Entropy
MDViT	Multi-domain Vision Transformer
ViT	Vision Transformers
NKT	Negative knowledge transfer
MKD	Mutual knowledge distillation
MHSA	Multi-head self-attention
MKD	Mutual Knowledge Distillation
DA	Domain Adapter
DSBN	Domain-Specific Batch Normalization
SE	Squeeze-and-Excitation
I2CVB	Initiative for Collaborative Computer Vision Benchmarking
TCIA	The Cancer Imaging Archive
FCN	Fully convolutional network
BN	Batch normalization
HCP	Human Connectome Project
ADNI	Alzheimer's Disease Neuroimaging Initiative
ABIDE	Autism Brain Imaging Data Exchange
IXI	Information eXtraction from Images
CSS	Continual semantic segmentation
B-MHA	Bidirectional multi-head attention
PaNN	Prior-aware Neural Network
PIPO-FAN	Pyramid Input Pyramid Output Feature Abstraction Network
DCNN	Deep convolutional neural network
MD	Multi-dataset
CPTM	Cross-patch transformer module
UMA-Net	Uncertainty-guided Multi-source Annotation Network
QAM	Quality Assessment Module
DFQ	Decoupled Feature Query
MiT-B3	Mix Transformer
KD	Knowledge distillation
DDA-GAN	Diverse data augmentation generative adversarial network
UDA	Unsupervised domain adaptation
DCNN	Deep convolutional neural networks
CMA	Cross-Modality Adaptation
RPA	Relation Prototype Awareness
IA	Inheritance Attention
MMWHS	Multi-Modality Atlases for Whole Heart Segmentation
FNN	Feedforward Neural Network
ELCFS	Episodic Learning in Continuous Frequency Space
FedAvg	Federated Averaging
CIIL	Cyclic Institutional Incremental Learning

ADMM	Alternating Direction Method of Multipliers
fPCA	Federated Principal Component Analysis
SAM	Segment Anything Model
NLP	Natural language processing
MedSAM	Segment Anything in Medical Images
SSM-SAM	Self-Sampling Meta SAM
FMAD	Flexible Mask Attention Decoder
Med-SA	Medical SAM Adapter
SD-Trans	Space-Depth Transpose
HyP-Adpt	Hyper-Prompting Adapter
MMDKD	Multi-Modal Decoupled Knowledge Distillation
SPPG	Self-Patch Prompt Generator
QDMD	Query-Decoupled Modality Decoder
MMDKD	Multi-Modal Decoupled Knowledge Distillation
HD	Hausdorff distance
BBox	Bounding box
AVD	Absolute volume Difference
RAVD	Relative Absolute Volume Difference
MSD	Mean Surface Distance
ASSD	Average symmetric surface distance
AUC	Area Under the curve
ASD	Average Surface Distance
T1	T1-weighted
T2	T2-weighted
FLAIR	Fluid-Attenuated Inversion Recovery
LASC	Left Atrial Segmentation Challenge
LiTS	Liver Tumor Segmentation
ADC	Apparent diffusion coefficient
PZ	Peripheral zone
TZ	Transition zone
GTVp	Gross Tumor Volumes
GTVn	Gross Tumor Volumes
HNSCC	Head-and-neck squamous cell carcinoma
AROI	Annotated Retinal OCT Images
CoNet	Coherent Network
IRF	Intraretinal Fluid
SRF	Subretinal Hyperreflective
PED	Retinal Pigment Epithelial Detachment
ILM	Internal Limiting Membrane
RPE	Retinal Pigment Epithelium
ME	Macular Edema
ASPP	Atrous Spatial Pyramid Pooling
SRHM	Subretinal Hyperreflective Material
IPL	Inner Plexiform Layer
INL	Inner Nuclear Layer
SNR	signal-to-noise ratio
RF	Random Forest
GSP	Graph-shortest path

SVDNA	Singular value decomposition
MUV	Medical University of Vienna
RUNMC	Radboud University Medical Centre
ROC	Receiver operating characteristics
SOTA	State-of-the-art
MMIS-Net	MultiModal Medical Image Segmentation Network
ViTDA_Net	Vision Transformer with domain adapters Network
CVD_Net	Convolutional Neural Network and Vision
Transformer with Domain-Specific Batch Normalization	
DSA	Domain-Specific Adapters
LoRA	Low-rank-based
EMC	Erasmus Medical Center
SCD	Skin Cancer Detection
HGJ	Hopital general juif, Montreal, Canada
CHUS	Centre hospitalier universitaire de Sherbooke, Sherbrooke, Canada
HMR	Hopital Maisonneuve-Rosemont, Montreal, Canada
CHUM	Centre hospitalier de l'Universite de Montreal, Montreal, Canada
CHUV	Centre Hospitalier Universitaire Vaudois, Switzerland
USZ	Universit¨atsSpital Zurich, Switzerland
CHB	Centre Henri Becquerel, Rouen, France
CHUP	Centre Hospitalier Universitaire de Poitiers, France
MDA	MD Anderson Cancer Center, Houston, Texas, USA
EHRs	Integration with Electronic Health Records

Chapter 1

Introduction

Medical image segmentation is a cornerstone of computer-aided diagnosis and health research. Manual segmentation, while essential, is often labor-intensive, requires significant expertise, prone to errors, and can introduce bias. To address these challenges, automated segmentation methods have emerged as a crucial focus of research, offering the potential for more efficient, accurate, and objective analysis of medical images. Deep learning methods have been successful in the segmentation and detection of diseases in medical imaging. However, most of these methods are trained and tested on images from the same source, modality, organ, or disease type, without fully exploring the synergistic potential of other datasets. This limitation leads to poor generalization when applied to new, unseen data, often encountered in real-world scenarios, due to a phenomenon known as the domain shift problem [25] where the distribution between the training (source) and testing (target) domains differs. This research aims to address this challenge through two key approaches: (i) improving the segmentation and generalization performance of specific designed models by capturing global contextual information at varying rates using Atrous Spatial Pyramid Pooling (ASPP) [35] and Squeeze-and-Excitation (SE) [74] blocks, and (ii) enhancing the segmentation and generalization performance of universal models by employing knowledge transfer techniques and domain adapters to effectively adapt and generalize to new domains and to mitigate Negative Knowledge Transfer (which occurs when knowledge learned from one domain negatively impacts another) [53] in diverse, multi-source datasets. The latter approach provides a promising solution to diversify the training set by incorporating multiple datasets from various sources, modalities, organs, and disease types, allowing a single model to learn from diverse examples. By leveraging the many small, annotated medical image datasets available in the public domain, this approach aims to explore the synergistic potential between datasets. However, simply merging data from different sources can lead to degraded performance due to Negative Knowledge Transfer. To address this, we incorporate knowledge adapters into the model architecture to capture domain-specific context from each domain while sharing common knowledge across all domains through a shared backbone. This approach mitigates the effects of negative knowledge transfer and enhances overall model performance.
This work introduces deep learning models that integrate content and domain adapters to tackle the challenges of domain shift and enhance model generalization. These advancements enable reliable disease diagnosis and effective monitoring of disease progression across diverse and unseen datasets.

1.1 Aim and Objectives

The aim of this thesis is to enhance segmentation and generalization performance in medical imaging by (i) capturing global contextual information through specific designed architectures that incorporate Atrous Spatial Pyramid Pooling (ASPP) and Squeeze-and-Excitation (SE) blocks, and (ii) employing domain adapters to effectively adapt and generalize to new domains and to mitigate negative knowledge transfer from diverse, multi-source data in a single generalizable architecture. The models developed in this research are designed to assist doctors in hospitals with diagnosing and monitoring the presence of diseases using medical images. The specific objectives of this PhD thesis are summarized as follows:

- **Enhancing Disease Detection, Segmentation and Generalization**:
  Develop a novel deep learning model to detect and segment diseases in medical images with high generalization performance, while handling high variability across diverse sources.
- **Capturing Global Contextual Information**:
  Develop a novel deep learning model to capture global contextual information in datasets with high variability, thereby improving the model's generalization performance.
- **Leveraging the Synergistic Potential of Combined Datasets**:
  Develop a novel, single, diverse model that leverages the synergistic potential of multiple small annotated datasets from diverse sources, modalities, organs, and disease types to improve the model's generalizability on new, unseen data.
- **Mitigating Negative Knowledge Transfer**:
  Develop a novel deep learning model to mitigate negative knowledge transfer in multi-source datasets using domain-specific adapters, thereby improving segmentation and generalization performance.

1.2 Contribution to Knowledge

A significant challenge in deep learning models is the domain shift problem [25], which arises when models are trained and tested on the same data source exhibit poor performance on new, unseen data typical of real-world scenarios. This performance degradation often stems from variations in image quality due to differences in vendor devices, scanning protocols, and the expertise of specialists capturing the images. Our contributions to addressing this challenge and enhancing disease segmentation and generalization performance in medical imaging are as follows:

- We propose a novel algorithm termed nnUNet_RASPP (nnU-Net with Residual and Atrous Spatial Pyramid Pooling). nnUNet_RASPP incorporates an Atrous Spatial Pyramid Pooling (ASPP) block to : (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. The ASPP block was positioned directly before the input layer and prior to downsampling to preserve contextual information. Also, we introduced residual connections in both the encoding and decoding paths to mitigate vanishing gradient problem within a convolutional neural network (CNN) backbone. Additionally, we conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed. The proposed nnUNet_RASPP was evaluated on the benchmark RETOUCH challenge dataset, which comprises of data from multiple sources acquired using three different device vendors. Experimental results on the hidden test set demonstrate that nnUNet_RASPP significantly outperforms state-of-the-art algorithms and large foundation models for medical image segmentation. nnUNet_RASPP also exhibited exceptional generalization performance on new, unseen data from diverse sources, surpassing state-of-the-art algorithms. Furthermore, we are the current winners of both the online and offline versions of the challenge.

- We further explore the potential of capturing global contextual features using Atrous Spatial Pyramid Pooling (ASPP) blocks to enhance segmentation and generalization performance. We propose a novel algorithm, Deep_ResUNet++, by integrating multiple ASPP and Squeeze-and-Excitation (SE) blocks at various locations within a convolutional neural network (CNN) backbone. This design captures global contextual information while dynamically adjusting the kernel size and network depth based on the input image size. Deep_ResUNet++ was evaluated on two public benchmark datasets: the Annotated Retinal OCT Images (AROI) and the Duke DME datasets, collected from patients with two distinct disease types. Experimental results demonstrate that Deep_ResUNet++ significantly outperformed state-of-the-art algorithms by a clear margin.

- The success of most deep learning models is often dependent on the availability of large datasets. However, in medical imaging, annotating images is a labor-intensive and time-consuming process which limits the size of available annotated datasets. Nevertheless, many small, publicly available annotated

datasets exist, spanning from different sources, organs, modalities, and disease types. We combine multiple datasets from these diverse domains to build a single and diverse model, leveraging the synergistic potential of one dataset on another to improve segmentation performance and generalization on unseen data. To accomplish this, we integrate knowledge transfer and domain-specific adapters to mitigate the effects of negative knowledge transfer within the backbones of two architectures: (i) a convolutional neural network (CNN) and (ii) a hybrid model combining CNN for feature extraction with a vision transformer (ViT) for long-range dependencies. This results in two novel architectures: MMIS-Net (MultiModal Medical Image Segmentation Network) and CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-Specific Batch Normalization). Both MMIS-Net and CVD_Net were evaluated on two groups of datasets. The first group includes 10 benchmark datasets covering 19 organs across two modalities, and the second group is the HECKTOR 2022 benchmark dataset, collected from nine medical centers around the world. Experimental results on the hidden test set show that MMIS-Net and CVD_Net outperformed state-of-the-art algorithms and large foundation models for medical image segmentation by a clear margin, while demonstrating high generalization capabilities on new, unseen data.

In this work we have developed deep learning models for the diagnosis and monitoring of diseases using medical images. The practical implications include automating the diagnostic and disease-monitoring processes, which are typically labor-intensive, time-consuming, and prone to errors. This automation will allow clinicians to focus on more complex tasks and can also serve as a decision-support tool, providing a valuable second opinion when making critical decisions.
Furthermore, early diagnosis and effective disease monitoring enable doctors to personalize and initiate treatment plans, such as tumor detection and therapy planning for cancer patients, improving patient care and reducing the socio-economic burden on both patients and healthcare systems.
This research also lays the groundwork for future studies, establishing a benchmark for result comparison and fostering further advancements in the field. By building on this work, future researchers can explore new techniques and applications in medical image segmentation.

1.3 Methodology

This research aims to enhance clinical outcomes and advance medical image analysis by addressing critical technical challenges such as capturing global contextual information, domain adaptation and model generalization. The work is centered on the development of innovative algorithms, with its scope defined by the following key areas:

Capturing Global Contextual Features: This research focuses on developing a robust and generalizable model. The work began by exploring the potential of capturing global contextual features to improve the generalization performance of specifically designed architectures. Initially, an Atrous Spatial Pyramid Pooling (ASPP) block was integrated at a single location just before the input layer within a convolutional neural network (CNN) backbone to: (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. Subsequently, the approach was extended by incorporating both ASPP and Squeeze-and-Excitation (SE) blocks at multiple locations within a CNN backbone.

Diverse Models: One way to improve the generalization performance of a model is by increasing the diversity of the training dataset. We developed single, diverse models by combining data from multiple sources, modalities, organs, and disease types, leveraging the synergistic potential of one dataset on another.

Similarity Fusion Block: Combining datasets from diverse sources, modalities, organs, and disease types often presents challenges with label inconsistencies. For instance, an anatomic structure or disease labeled in one dataset may not be labeled in the same organ in another dataset. To address this, we introduce Similarity Fusion, a novel technique designed to capture cross-dimensional dependencies in feature maps while effectively managing datasets with inconsistent labels.

Domain Adaptation: Naively combining data from multiple sources can improve performance on one dataset but lead to degradation on another due to negative knowledge transfer, ultimately resulting in overall poor model performance. In this work, we utilize domain adapters to effectively extract domain-specific information while sharing common features across all domains through a shared backbone. This approach mitigates the effects of negative knowledge transfer, thereby enhancing the model's generalizability.

Benchmarks and Hidden Test Sets: The algorithms presented in this work were experimented on benchmark datasets, and for fair comparison to other state-of-the-art architectures, they were evaluated on hidden test sets (where the raw data is publicly available, but the annotated/ground truth data is hidden). The results are published online on the respective challenge websites.

Hybrid Model: The two predominant networks in deep learning are Convolutional Neural Networks (CNN) and Vision Transformers (ViT). This work presents CNN-based models and a hybrid model that combines CNN for feature extraction with ViT for capturing long-range dependencies.

Literature Review and Public Resources: In Chapter 2, we provide a comprehensive review of the literature, classifying previous work into 6 main areas: Specific Models, Domain Adaptation, Universal Models, Federated Learning, Fine-tuning, and Foundation Models. Additionally, we include links to GitHub repositories where

authors have shared their source code publicly, along with links and descriptions of large, publicly available annotated multi-modal medical image datasets.

This research lays the foundation for future advancements in medical image segmentation, establishing a benchmark for continued innovation in the field.

1.4 Data Collection

The datasets used in this work are publicly available benchmarks, with all descriptions and links provided.

1.5 Structure of The Thesis

In this section, we outline the structure and organization of this work. The thesis is structured as follows:

- **Introduction**:
  Chapter 1 presents a summary of the thesis, the aim and objectives, the contribution to knowledge, scope, data collection, and road map.

- **Literature Review**:
  Chapter 2 provides an in-depth review of the existing literature relevant to this work. The reviews are categorized into 6 main areas: specific models, domain adaptation, universal model, federated learning, fine-tuning, and foundation models. Additionally, we present a summary of several large collections of public available annotated medical image datasets from various sources.

- **Enhancing Retinal Disease Detection, Segmentation, and Generalization with an ASPP Block and Residual Connections Across Diverse Data Sources**:
  Chapter 3 introduces nnUNet_RASPP (nnU-Net with Residual and Atrous Spatial Pyramid Pooling), a novel approach for disease detection, segmentation, and generalization in retinal optical coherence tomography (OCT) images from multiple sources.

- **Dynamic Network for Global Context-Aware Disease Segmentation in Retinal Images Using Multiple ASPP and SE Blocks** :
  Chapter 4 introduces a novel algorithm, Deep_ResUNet++, for disease and layer segmentation in retinal images.

- **Enhancing Medical Image Segmentation Through Knowledge Transfer with Domain-Specific Adapters Across Diverse Data Sources**:
  Chapter 5 presents novel approaches to enhance model performance and generalizability on new, unseen data by creating a single, diverse, and generalizable model that combines data from multiple sources, modalities, organs, and disease types. The chapter proposes two novel methods using knowledge transfer and domain-specific adapters: one integrates domain-specific adapters within a convolutional neural network (CNN) backbone, and the other combines a

CNN for feature extraction with a Vision Transformer (ViT) to capture long-range dependencies in the backbone.

- **Discussion and Conclusion**:
  Chapter 6 is a discussion chapter that begins with an overview of the approaches, contributions, and results from Chapters 3, 4, and 5. It then provides a detailed assessment of each approach in relation to the research objectives outlined in Section 1.1. Additionally, the chapter addresses the implications and limitations of the methods presented and potential directions for future research.

Chapter 2

Literature Review

Deep learning has revolutionized computer-aided detection and diagnosis (CAD) in medical image analysis. However, the performance of these models often deteriorates when faced with data from different sources, a phenomenon known as domain shift, where models often perform poorly given out-of-distribution examples. To address this challenge, knowledge adapters and model generalizability have emerged as promising solutions. This chapter presents a comprehensive review of recent advancements in deep learning methods for medical image analysis, focusing on tasks such as disease diagnosis, lesion and organ detection, and abnormality detection from diverse data sources tackling the problem of domain shift. We categorize existing methods into 6 main areas: Specific Models, Domain Adaptation Models, Universal Models, Federated Learning Models, Fine-Tuning Models, and Foundation Models, and discuss their effectiveness in handling heterogeneous data sources. Furthermore, we highlight relevant benchmark datasets and identify key challenges and future research directions in this rapidly evolving field. This survey aims to provide researchers with a solid understanding of the current state-of-the-art and inspire innovative approaches to improve the generalizability of deep learning models for medical image analysis.

2.1 Introduction

Medical segmentation tasks span a broad spectrum of imaging modalities, including optical coherence tomography (OCT), computed tomography (CT), magnetic resonance imaging (MRI), and X-rays. These modalities are applied to various anatomical structures, such as the abdomen, chest, brain, retina, head, and even individual cells, to identify conditions like cancerous cells, tumors, fluid accumulations, organ abnormalities, and more. This diversity has led to the development of numerous segmentation tools, each typically designed to address a specific task or a small set of related tasks. In recent years, deep learning has been widely applied to medical image segmentation, classification, and analysis, often under the assumption that the training and test datasets share the same data distribution [185]. However, this assumption frequently does not hold in practice (real world scenarios). Research has shown that test error typically increases in proportion to the distributional differences between training and test datasets [16], [182], a challenge known as the “domain shift“ problem [25]. Therefore, addressing domain shift is critical for the effective application of deep learning methods in medical image

segmentation, classification, and analysis.

Numerous small annotated datasets from diverse sources, organs, modalities, and disease types, collected using different vendor devices, are available online such as [21], [133], [71], [8], [130], [40], [120] and many more. An intuitive approach is to build a single robust model by combining datasets from these various sources. However, the domain shift problem persists across different medical image datasets due to variations in imaging modalities, disease types, scanning parameters, expertise levels, subject cohorts, and other factors. To address these challenges and improve model generalizability, domain knowledge adapters have emerged as a promising solution [65]. Researchers have increasingly focused on utilizing these small annotated datasets from the public domain to tackle various tasks in medical image segmentation, classification, and analysis.

A summary of previous research including references, year, methodology, target organ, image dimensions, and metrics is provided across several tables: Table 2.1 for specific models, Table 2.3 focuses on domain adaptation approaches, Table 2.2 highlights universal models, Table 2.4 covers federated learning methods, Table 2.6 presents fine-tuning models, and Table 2.5 details large foundation models. Table 2.7 provides public available large datasets from diverse sources. Finally, Table 2.8 provides a summary of the authors who have made their code publicly available, along with their corresponding GitHub links.

The rest of this chapter is organized as follows: A brief overview of key specific model approaches in medical image segmentation from diverse data sources for medical image analysis categorized into different sections. Section 2.2 presents specifically designed models, while universal models are discussed in Section 2.3. Domain adaptation techniques are covered in Section 2.4, and federated learning approaches are explored in Section 2.5. Foundation models are discussed in Section 2.6, followed by fine-tuning techniques in Section 2.7. Section 2.8 provides a summary of large benchmark medical image datasets from diverse sources. Finally, a summary of the literature and identified gaps is presented in Section 2.9.

2.2 Specific Designed Model Approaches for Medical Image Segmentation

In the diagnosis and segmentation of diseases in medical images using deep learning, the three most widely adopted base architectures are convolutional neural networks (CNN), U-Net, and vision transformers (ViTs). To enhance segmentation performance and generalization across data from diverse sources, many researchers have employed CNN, U-Net, ViT, or a combination of these backbones. Before exploring generalizable models, we will first review some of the recent CNN, U-Net and ViT based task specific approaches.

The ReLayNet architecture a CNN-based network was introduced in [160] for the segmentation of layers and fluids in OCT images. The ReLayNet enhanced the kernel shape to match the shape of the input image. The algorithm was experimented on the publicly available Duke dataset, which comprises of 110 annotated B-scans (divided into 10 classes: 1 background, 8 layers, and 1 fluid) acquired from ten patients with Diabetic Macular Edema (DME)[40]. Another CNN-based approach was presented in [118], which focused on retinal fluid segmentation and detection in OCT images. This framework is specifically designed to identify and segment three types of retinal fluids: Intraretinal Fluid (IRF), Subretinal Hyperreflective Material (SRF), and Retinal Pigment Epithelial Detachment (PED). The method is comprised of three main stages which are a pre-processing layer, a feature extraction layer, and a classification layer. Similarly, another CNN-based approach for the automatic segmentation of nine retinal layer boundaries in OCT images of patients with dry Age-related macular degeneration (AM) was presented in [54]. The authors used a regular Convolutional Neural Network (CNN) to extract features of the layer boundaries from the input image and classified them into nine classes, each representing one of the layer boundaries. Additionally, they applied a graph search method to further classify the extracted features into ten classes using probabilistic methods, aimed at eliminating misclassified features. Another CNN approach was reported in [101] to segment fluid from 1,289 OCT images of patients with Macular Edema (ME).

Since the introduction of the 2D U-Net [157], a convolutional network for biomedical image segmentation, in 2015, it has become the standard backbone for numerous medical image segmentation tasks. U-Net's architecture is characterized by its U-shape, consisting of an encoder path for capturing contextual features and a decoder path for precise pixel localization. These two paths are connected by a bottleneck that ensures a smooth transition between the encoder and decoder. Both the encoder and decoder, along with the bottleneck, are composed of convolutional blocks. At the end of the decoder path, there is a classification layer which assigns each pixel to one of the segmentation classes. Some of the derivatives of the 2D U-Net includes: The ResUNet architecture [216], which was originally developed for road image extraction. The ResUNet incorporated residual blocks into the U-Net back backbone. An extension of ResUNet for for medical image segmentation for domain-specifi task specifically targeting colonoscopic images is presented in [88] termed ResUNet++. The ResUNet++ was experimented on the Kvasir-SEG and CVC-612 datasets outperforming other model. Moving on, the Md-Unet, a multi-input dilated U-Net architecture designed for bladder cancer segmentation, was presented in [64]. The Md-Unet, modified standard convolution, by introducing a hyper-parameter called

dilated rate, which referred to the number of kernel intervals. The algorithm was evaluated on a private bladder cancer dataset from Yunnan University, which contains 768 lesion images. Experimental results indicate that the Md-Unet achieved performance comparable to other state-of-the-art algorithms. Building from the succcess of the 2D U-Net researchers at Moorfields Eye Hospital NHS Foundation Trust London and DeepMind Health extended the standard 2D U-Net to present a 3D U-Net in [43] for the diagnosis and referral in retinal disease. The architecture consisted of two parts: a segmentation model and a classification model. The algorithm was trained on 14,884 OCT scan volumes obtained from 7,621 patients. Moving on, USE-Net, which enhanced the 2D U-Net architecture with Squeeze-and-Excitation (SE) blocks for prostate zonal segmentation on multi-institutional MRI datasets, was presented in [161]. The integration of SE blocks aimed to improve segmentation accuracy by modeling channel-wise dependencies in convolutional features, applied after each convolutional layer in both the encoding and decoding paths of U-Net. The model was trained on individual and multi-site prostate MRI datasets collected from different institutions. Experimental results indicated that USE-Net significantly outperformed state-of-the-art algorithms in prostate zonal segmentation across heterogeneous datasets. Inspired by the success of U-Net for medical image segmentation, the nnU-Net ("no new-Net"), a self-configuring method for deep learning-based biomedical image segmentation, was introduced in [84]. The nnU-Net is based on the U-Net architecture, but instead of relying on manual parameter tuning "trying an error" methods, nnU-Net proposed a self-parametrizing pipeline. The pipeline generates a "data fingerprint" by analyzing the training data and uses key dataset properties, such as modality, shape, and spacing, to automatically configure key model parameters like network topology, image resampling methods, input patch sizes, and kernel sizes, based on graphics processing unit (GPU) availability and hardware constraints. During training, data augmentation is applied on the fly, and after training, the framework determines "empirical parameters" for post-processing. Certain parameters, such as the loss function, remain fixed throughout training, with the framework using a combination of Cross Entropy and Dice loss functions. The framework was evaluated on 11 international biomedical image segmentation challenges, consisting of 23 different datasets and 53 segmentation tasks, achieving first place in 33 out of the 53 tasks.

Since the introduction of nnU-Net, several of its variants have been proposed. In [126], residual, dense, and inception blocks were integrated into the network, and the approach was evaluated on eight datasets consisting of 20 target anatomical structures. Advanced architectural variations of the network were explored in [127] and evaluated on eight medical imaging datasets covering 20 anatomical regions. In the Multi-Center Fetal Brain Tissue Annotation (FeTA) 2022 challenge [146], the standard nnU-Net or its variants were used by the top five teams. This benchmark involved fetal brain MRI data acquired from four different centers. A comparative analysis and performance evaluation of the nnU-Net was presented in [78], and experimented on multiple data sources. A study titled nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation, was proposed in [83], providing a comprehensive evaluation of the nnU-Net variants. The study was experimented on six datasets, emphasizing the importance of rigorous validation in 3D medical image segmentation.

Following the recent success of Transformers in natural language processing

(NLP) such as [2], [183], and [205], researchers have sought to replicate their effectiveness in deep learning by using Vision Transformers (ViTs). Some of the ViTs models used for medical image segmentation will be briefly disscused as follows: SegFormer3D, an efficient transformer for 3D medical image segmentation, was introduced in [149]. It is a Vision Transformer (ViT) based hierarchical model that computes attention across multiscale volumetric features. The architecture featured an all-MLP (multilayer perceptron) decoder that combined local and global attention features to generate precise segmentation masks. SegFormer3D was evaluated on three benchmark datasets, achieving results competitive with state-of-the-art algorithms. Similarly, Swin UNETR, another transformer model for medical image segmentation, was introduced in [66] and was specifically designed for the segmentation of 3D brain tumors. Swin UNETR features a hierarchical transformer encoder for feature map extraction, a self-attention mechanism for skip connections, and a CNN decoder for up-sampling. The model employed the Dice loss function and was evaluated on the BraTS 2021 dataset [12], achieving results comparable to state-of-the-art architectures. Building on the success of Swin UNETR, the authors proposed an enhanced variant called UNETR in [67], which incorporated a modified loss function combining soft Dice loss and cross-entropy loss. UNETR was evaluated on the BTCV [100] and MSD [8] datasets, demonstrating performance comparable to state-of-the-arts algorithms.

Other researchers have explored hybrid approaches that combine Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), some of which are briefly outlined as follows: The TransClaw U-Net, a hybrid model combining the U-Net and Transformer architectures, was introduced in [30]. The TransClaw U-Net consists of an encoder and a decoder path. The encoder includes convolutional blocks for extracting shallow spatial features and Transformer blocks for capturing global features, while the decoder path uses convolutional blocks for pixel localization. The model was evaluated on the Synapse Multi-Organ Segmentation dataset[177] and achieved performance comparable to state-of-the-art algorithms. Similarly, SwinBTS, another hybrid model integrating CNNs and Vision Transformers, was proposed in [91]. SwinBTS used Transformer modules for feature extraction and convolutional operations for up and down sampling. It was evaluated on three brain segmentation datasets: BraTS 2019, BraTS 2020, and BraTS 2021, achieving performance comparable to state-of-the-art architectures. Moving on, another notable hybrid model is the nnFormer, a volumetric medical image segmentation model introduced in [220]. This approach combines CNNs and 3D Transformers for effective disease segmentation in medical images. The nnFormer leverages the self-parameterization, pre-processing, and post-processing capabilities of nnU-Net. It was evaluated on three benchmark datasets.

One limitation of specific designed algorithms is their task-specific nature. They perform well on a particular task, organ, or disease type but may struggle or perform poorly when applied to other tasks.

Reference	Year	Backbone	Organ	Modalities	Dimensions	Metrics
U-Net[157]	2015	U-Net	Multiple	Multiple	2D	IoU
[54]	2017	CNN	Eye	OCT	2D	MD
ReLayNet [160]	2017	U-Net	Eye	OCT	2D	DS
[101]	2017	CNN	Eye	OCT	2D	DS
[43]	2018	U-Net	Multiple	OCT	3D	DS
ResUNet++ [88]	2019	U-Net	Colon	Endoscopic	2D	DS/IoU
USE-Net [161]	2019	U-Net	Prostate	MRI	2D	DS
Swin UNETR [66]	2021	ViT	Brain	MRI	3D	DS
Md-Unet [64]	2021	U-Net	Bladder	MRI	3D	DS/IoU
nnU-Net [84]	2021	U-Net	Multiple	Multiple	2D/3D	DS
Transclaw [30]	2021	CNN/ViT	Multiple	CT	3D	DS/HD
nnFormer [220]	2021	CNN/ViT	Multiple	MRI/CT	3D	DS/HD
UNETR in [67]	2022	ViT	Multiple	MRI/CT	3D	DS
SwinBTS [91]	2022	CNN/ViT	Brain	MRI	3D	DS/HD
[126]	2022	U-Net	Multiple	MRI	3D	DS
[127]	2023	U-Net	Multiple	Multiple	3D	DS
[146]	2024	U-Net	Multiple	MRI	3D	DS
[78]	2024	U-Net	Multiple	Endoscopic	3D	DS
[83]	2024	U-Net	Multiple	Multiple	3D	DS
SegFormer3D [149]	2024	ViT	Brain	MRI/CT	3D	DS

Table 2.1: A summary of previous work on specific models, listed in order of year of publication, including the references, year, backbone, organ, modalities, image dimensions, and evaluation metrics: Intersection over Union (IoU), Mean Difference (MD), Dice Score (DS), Hausdorff distance (HD).

2.3 Universal Model Approaches for Medical Image Segmentation

Another way to reduce the domain shift between training and testing datasets, thereby enhancing the model's generalizability, is by increasing the diversity of the training data through the integration of data from multiple diverse sources to build a single universal model. In this section, we will briefly review some of these approaches based on their architectural backbone, in the following order: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and a combinations of both in a hybrid model.

The generalizable multi-site training and testing of deep neural networks, aimed at addressing high image variability and intensity differences across datasets collected from various sites using different scanners, was presented in [141]. This approach modified U-Net to build a single segmentation model spanning data from multiple sites. The U-Net architecture was enhanced by reducing the number of max-pooling operations from four to three, resulting in a network with 18 convolutional layers and 7,696,256 trainable parameters. A patch-based training approach was adopted, using patches of 128×128 pixels. The algorithm was tested on 600 magnetic resonance (MR) prostate gland segmentation images from two different sites.

Another CNN-based architecture was presented in [48] which built a single multi-class segmentation model by combining several single-class datasets. This approach proposed a unified and efficient framework for robust multi-class segmentation by combining single-class datasets and conditioning a convolutional network for segmentation tasks. The algorithm modified the U-Net backbone by incorporating a conditioning module, making the model fully convolutional, simple, and efficient, thereby avoiding performance overhead. Unlike other approaches that trained separate models for each class, this method inferred segmentations and captured relationships among multiple classes from single-class datasets, enabling a single model to generate segmentations for all classes. The algorithm was evaluated on three public datasets of abdominal CT volumes.

Similarly, another CNN-based architecture was presented in [171], which used marginal loss and exclusion loss functions for multi-organ segmentation to train a single multi-organ segmentation network using multiple datasets. This approach modified the nnU-Net framework by incorporating these two loss functions. In most annotated medical image datasets, only regions of interest (ROIs) are labeled, while other areas are classified as background, often referred to as partially labeled datasets. Combining data from multiple partially labeled datasets can create conflicts, where a region annotated in one dataset might be classified as background in another. To address this, [171] proposed two loss functions: (i) the marginal loss, which merges all unlabeled organs with the background by assuming a marginal probability for combining the background label and can be incorporated into cross-entropy (CE) loss Dice loss, and (ii) the exclusion loss, which enforces exclusivity by ensuring a one-to-one mapping between each labeled pixel and its corresponding label. The algorithm was evaluated on five benchmark organ segmentation datasets from various sources, demonstrating its effectiveness.

Moving on another CNN-based architecture was proposed in [90]. This framework introduced a continual semantic segmentation (CSS) approach that used a

CNN encoder-decoder network architecture to sequentially learn a single multi-organ segmentation model from multiple partially labeled datasets, one at a time. After training on a specific dataset, the encoder was frozen, and decoders for other datasets learned from the extracted feature maps. The outputs of these decoders were then merged to create a single model, contributing to a universal organ segmentation system. The model was initially trained on a specific dataset, optimizing both the shared encoder and its associated decoder. Once trained, the dataset was no longer accessible, and subsequent stages focus on training decoders for new datasets while keeping the general encoder frozen. The predictions from all decoders were combined, resulting in a model capable of multi-organ segmentation. This cycle was repeated for all datasets. The framework was evaluated on 103 anatomical structures from one public and three private partially multi-organ datasets.

Another CNN-based approach aimed to address the issue of partially labeled datasets was presented in the multi-organ segmentation via co-training weight-averaged models from few-organ datasets [79]. This method introduced a unified model that integrated data from multiple sources to tackle the challenges of limited generalization and noisy pseudo labels. The method proposed co-training weight-averaged models to train a multi-organ segmentation network from datasets with annotations for only a few organs. The approach involved collaboratively training of two networks, which thought each other about unannotated organs. During training, two networks were used, but only one was required for inference, ensuring no additional computational or memory overhead at the inference stage. The framework was evaluated on three publicly available single-organ datasets.

The Pyramid Input Pyramid Output Feature Abstraction Network (PIPO-FAN), a CNN-based approach for multi-organ segmentation on partially labeled datasets using multi-scale feature abstraction, was introduced in [55]. PIPO-FAN constructed a single segmentation network by integrating pyramid inputs and feature analysis into a U-Net backbone to achieve multi-scale feature abstraction. It employed an equal convolutional depth mechanism to merge features from different scales, a deep supervision mechanism to refine outputs at various scales, and an adaptive weighting layer to automatically fuse the outputs. The architecture consists of three key components which are: (i) a pyramid-input and pyramid-output network that condensed multi-scale features and reduced the semantic gaps between features from different scales, (ii) an image context-based adaptive weighting layer that fused segmentation features across multiple scales, and (iii) a target-adaptive loss integrated with a unified training strategy to enable segmentation across multiple partially labeled datasets using a single model. PIPO-FAN was evaluated on four publicly available datasets.

Tgnet, a task-guided network architecture for multi-organ and tumor segmentation from partially labeled datasets was presented in [194]. Tgnet built a single segmentation model by efficiently learned task-specific features while preventing the mixing of representations from different organs and tumors across tasks. The approach enhanced the standard U-Net architecture by incorporating a modified residual block and attention module to fuse image features with task-encoding constraints in a task-guided manner. Tgnet claimed it's design effectively suppressed irrelevant features and emphasized features relevant to each specific segmentation task. The algorithm has an encoder-decoder structure, with a task-guided attention module placed in the skip connections, using a global average pooling (GAP) mod-

ule to capture global contextual information, task-guided residual blocks to mitigate overfitting, task-encoding concatenation, and a convolutional layer with a sigmoid activation function. The network was evaluated on seven partially labeled organ and tumor datasets.

Another CNN-based approach to addressing the challenge of partially labeled datasets was presented in DoDNet [209]. DoDNet constructed a sigle segmentation model by learning to segment multi-organ and tumors from multiple partially labeled datasets. The authors proposed a dynamic on-demand network (DoDNet), which can be trained on partially labeled datasets for multi-organ and tumor segmentation. DoDNet featured a U-Net like encoder-decoder architecture with a single, dynamic head capable of performing tasks typically handled by multiple networks or a multi-head network. The kernels in the dynamic head were generated adaptively by a controller, conditioned on the input image and task. For each segmentation tasK, task-specific priors guide the controller to generate kernels dynamic head. The algorithm was evaluated across seven organ and tumor segmentation benchmarks.

Moving on, the Omni-Seg, a unified dynamic network for multi-label renal pathology image segmentation using partially labeled data, was introduced in [44]. Inspired by DoDNet [209], it extends the U-Net architecture with three key components which are: (i) A Dynamic Multi-Label Modeling, which encoded class-specific information for different tissue types into an m-dimensional one-hot vector. This vector was merged with feature embeddings at the deepest layer of the residual U-Net helping the network to learn domain-specific information for each class. (ii) A Dynamic Head Mapping, which is a binary segmentation network that employed dynamic filters to target specific tissue types. It optimized feature vectors and class-aware vectors, using these to guide a lightweight dynamic head comprising three layers (two with eight channels, and a final layer with two channels). And (iii) Residual connections that featured multiple encoder-decoder blocks, arranged in a pyramid structure. The decoder upscaled feature maps, combining them with low-level features from the encoder using residual blocks, leading to refined high-level segmentation outputs. The method was experimented on 1,751 regions of interest (ROIs) from 459 whole slide imaging scans from 125 patients.

The Diverse Data Augmentation Generative Adversarial Network (DDA-GAN), a CNN-based adversarial network approach was introduced in [38] for learning image segmentation using cross-modality annotations. The approach trained a single segmentation model for an unannotated target domain by utilizing information from an annotated source domains. This was achieved by generating diverse augmented data for the target domain through a one-to-many source-to-target translation technique. Key components of the framework includes: (i) An S-feature Encoders that extracted structural information from both source and target domains. (ii) A-feature encoders, that captured appearance information, using global average pooling and fully connected layers that enhanced non-linear mapping and eliminated positional variations. (iii) A decoder to generate images by fusing S-features and A-features through residual and decoding blocks. (iv) A segmenter to annotate images based on their S-features, using a similar architecture as the decoders. (v) An image discriminators, that distinguished between real and generated images in both source and target domains. And (vi) A feature discriminator to identify whether S-features come from the source or target domain. DDA-GAN enabled effective single model segmentation from diverse data sources by leveraging both structural and appear-

ance features across different modalities. The framework was evaluated on two datasets.

The Uncertainty-guided Multi-source Annotation Network (UMA-Net), a CNN-based approach, was presented in [4] for developing a robust, universal medical image segmentation model from multiple data sources. Built on the U-Net backbone, UMA-Net integrated two key components that guided the training process using uncertainty estimation at both the pixel and image levels. The first component, the Annotation Uncertainty Estimation Module (AUEM), estimated pixel-wise uncertainty for each annotation dataset, enabling the network to focus on reliable pixels by applying a weighted segmentation loss. The second component, the Quality Assessment Module (QAM), evaluated the quality of the image using the pixel-wise uncertainties assessed by the AUEM. UMA-Net was evaluated on three datasets, demonstrating its capability to handle multi-source annotations effectively.

Transitioning from CNN to Vision Transformer (ViT) backbone models, the MedFormer, a unified and data-scalable transformer model for medical image segmentation, was introduced in [61]. MedFormer was designed for 3D medical image segmentation, combining data from diverse sources, modalities, organs, and disease types to create a universal model. It incorporated depth-wise separable convolution within transformer blocks to embed desirable inductive biases. A key innovation of MedFormer was the Bidirectional Multi-Head Attention (B-MHA) mechanism, which reduced redundant tokens through low-rank projection, effectively lowering the quadratic complexity of conventional self-attention to linear. This enabled efficient modeling of long-range relationships, capturing global interactions in high-resolution token maps and improving detailed boundary modeling. Other notable components of MedFormer included a Vision Transformer (ViT) for capturing long-range dependencies, an efficient attention mechanism that reduced token numbers via subsampling layers, a global multi-scale semantic fusion map for multi-scale feature integration, and a convolutional inductive bias to address the loss of local structure information inherent in transformers. During training, each pixel was treated as a token, and the token map was flattened into a sequence for processing by the transformer block. The model was validated on eight widely used public datasets across various modalities and target structures, demonstrating its effectiveness.

The Mix Transformer (MiT-B3), a transformer based semantic segmentation model originally designed for natural images, by incorporating cross-attention and self-attention mechanisms to decouple feature queries, thereby improving generalization was presented in [198]. The Decoupled Feature Query (DFQ) framework, an enhanced variant of the MiT-B3 for medical image segmentation was presented in [17] to build a single model aimed at improving domain generalization in medical image segmentation across multiple source and unseen target domains. The framework focused on learning generalized representations through two key components: (i) Learning from Decoupled Feature Queries, which leveraged long-range dependencies in the self-attention mechanism to generate high-level feature queries from deep features, while the corresponding keys and values were derived from shallow features; and (ii) Decoding Generalized Representations, which fused the generalized features using a linear layer, subsequently processed by the segmentation head to produce final predictions. During training, DFQ used feed-forward layers and normalization within Transformer blocks to ensure consistent shallow feature representations across domains, facilitating the learning of robust and generalized features. The al-

gorithm was evaluated on benchmark datasets for fundus and prostate segmentation, demonstrating its effectiveness.

The CLIP-Driven Universal Model for organ segmentation and tumor detection was presented in [110]. Drawing inspiration from the success of Transformer models in natural language processing (NLP) such as [2], [183], and [205]. The model integrated text embeddings with voxel-level semantic segmentation, enabling segmentation across various datasets, organs, tumors, tasks, and imaging modalities. The framework consisted of two branches: a text branch and a vision branch. The text branch generated CLIP embeddings for each organ and tumor using specialized medical prompts, while the vision branch used both the images and the embeddings to predict segmentation masks. To address label inconsistency, CLIP-driven Universal Model incorporated text embeddings and used a masked back-propagation mechanism with binary segmentation masks. During training, the text branch produced CLIP embeddings based on medical prompts, which were concatenated with global image features and passed into a multi-layer perceptron (MLP). The vision branch processed CT scans by applying isotropic spacing and uniform intensity scaling to reduce domain gaps between datasets. The scans were passed through an encoder to extract feature maps, with the final layer generating predictions for each class in a one-vs-all manner. The CLIP-Driven Universal Model was evaluated on 14 public datasets comprising 3,410 CT scans.

Moving on to the hybrid models, other researchers have explored the potentials of a single hybrid models by combining CNN and ViT together. A combination of a convolutional neural network and a Transformer (CoTr), in an encoder-decoder structure for 3D medical image segmentation, was presented in [199]. The authors enhanced the nnU-Net framework by integrating a convolutional neural network (CNN) to extract feature representations and an efficient deformable Transformer (DeTrans) to model long-range dependencies on the extracted feature maps. CoTr consisted of a CNN encoder, a DeTrans encoder, and a CNN decoder. During training, the input images were flattened into a 1D vector, and a small set of key positions in the image were passed through the deformable self-attention mechanism. This approach reduced computational and spatial complexities, allowing for multi-scale and high-resolution feature map processing. The Transformer dynamically adjusted the receptive field based on the input content, enabling effective convolutional operations for modeling long-range dependencies. The algorithm was evaluated on the Multi-Atlas Labeling Beyond the Cranial Vault (BCV) challenge dataset [203].

Similarly, the TransUNet, another unified hybrid model combining a convolutional neural network (CNN) and a Transformer, was presented in [34]. Within an encoder-decoder structure, the TransUNet used a CNN encoder to extract global context feature maps, which were fed into a Transformer encoder for patch-based tokenization. The decoder upsampled the encoded features and combined them with high-resolution CNN feature maps to enhance pixel localization. The algorithm was experimented on the Synapse multi-organ segmentation [177] and ACDC [11] datasets. Other Variants of the TransUNet, have also been proposed for a single model for medical image segmentation using multi-source datasets, which will be reviewed as follows. A versatile medical image segmentation approach, leveraging model self-disambiguation to learn from multi-source datasets, was presented in [37]. Built on the TransUNet backbone, this method employed a hierarchical sampling technique to generate training examples from multi-source and multi-modality

datasets with ambiguous annotations. A 3D variant of TransUNet, referred to as 3D TransUNet, was introduced in [33]. This model extracted per-voxel feature representations from input volumes using the TransUNet backbone, which were processed by a segmentation head to produce multi-channel predictions. The 3D TransUNet incorporated prior knowledge through model self-disambiguation, encouraging confident and informative predictions, and used a hierarchical sampling approach to handle variations in imaging modalities, equipment, protocols, and patient demographics. During training, the model employed a multi-stage sampling strategy. It began by selecting images based on the type of anatomical structure to narrow down the eligible images. Next, images were sampled according to their modality from the refined subset, followed by sampling based on their dataset of origin to ensure equitable representation from diverse sources. Finally, an image is chosen from the selected dataset for training. The framework was validated on 2,960 volumetric images from eight multi-modal sources, including seven public datasets focused on abdominal structures, demonstrating its effectiveness in multi-source medical image segmentation.

Another hybrid approach to develop a universal model for medical image segmentation across diverse data sources was presented in [111]. A key innovation of the architecture was the integration of a cross-patch transformer module (CPTM) into the nnU-Net framework. The CPTM enhances segmentation performance by fusing information from adjacent image feature patches, by expanding the receptive field, and improving long-range context modeling, which is crucial for accurate segmentation. During training, a shared encoder extracted features from each patch, which are then flattened into 1D vectors. The flattened features were processed through multiple CPTMs for information fusion. The information from adjacent patches were merged into a central patch, which was subsequently decoded for segmentation prediction. The CPTM used two types of transformer blocks: one that fused global information within a single patch and another that fused information between adjacent patches. The model was trained on 33 anatomical categories across 7 partially-labeled datasets, encompassing around 2,800 volumes from three categories (3 pelvic bones, 5 abdominal organs, and 25 vertebrae).

Both CNN and ViT have demonstrated great success in medical image segmentation. But one notable limitation of CNN and ViT based models is their reliance on large training datasets. As demonstrated in [51], CNNs outperform ViTs when trained on datasets of comparable size. This is expected, as Transformers lack certain inductive biases inherent to CNNs, such as translation equivariance and locality, which enable CNNs to generalize better with limited data. However, when trained on very large datasets, ViTs surpass CNNs in performance, leveraging their ability to model long-range dependencies effectively. Another notable limitation is the imbalance in dataset sizes and modalities. Since most datasets are sourced online and collected from various medical centers, there is significant variability in their sizes, leading to imbalances in both dataset size and modalities.

Reference	Year	Backbone	Organ	Modalities	Dimensions	Metrics
UMA-Net [4]	2017	CNN	Eye	Fundus	2D	Accuracy
[141]	2019	CNN	Multiple	MRI	3D	DS
[48]	2019	CNN	Multiple	CT	3D	DS
PIPO-FAN [55]	2020	CNN	Heart	CT	3D	DS
[79]	2020	CNN	Multiple	CT	3D	DS/HD
TransUNet [34]	2021	ViT	Multiple	MRI/CT	3D	DS/HD
[171]	2021	CNN	Multiple	CT	3D	DS/HD
DoDNet [209]	2021	CNN	Multiple	CT	3D	DS/HD
DDA-GAN [38]	2021	CT	Multiple	MRI/CT	3D	DS
Tgnet[194]	2022	CNN	Multiple	CT	3D	DS/HD
MedFormer [61]	2022	ViT	Multiple	MRI/CT	2D/3D	DS/HD
CoTr[199]	2021	CNN/ViT	Multiple	CT	3D	DS
CPTM [111]	2022	CNN/ViT	Multiple	CT	3D	DS/HD
CLIP-Driven [110]	2023	ViT	Multiple	CT	3D	DS
[90]	2023	CNN	whole-body	CT	3D	DS/HD
3D TransUNet [37]	2024	ViT	Multiple	MRI/CT	3D	DS
[17]	2024	ViT	Eye	Fundus	2D	DS/ASD

Table 2.2: A summary of previous work on universal models, listed in order of year of publication, including the references, year, method, organ, image dimensions, and evaluation metrics: Relative Absolute Volume Difference (RAVD), Accuracy, Dice Score (DS), Hausdorff Distance (HD), and Average Surface Distance (ASD).

2.4 Domain Adaptation Approaches for Medical Image Segmentation

Most deep learning methods suffer from the domain shift problem, which occurs when a model trained on one dataset (source domain) performs poorly when tested on a new unseen dataset (target domain). This issue occurs because of the differences in the data distributions between the source and target domains. One approach to overcoming this challenge is to use Domain Adaptation (DA)techniques during training. DA techniques help the model to capture domain-specific features while sharing common features within a universal network, thereby enhancing the model's generalization ability. An illustration to tackle domain shift is shown in Figure 2.1.

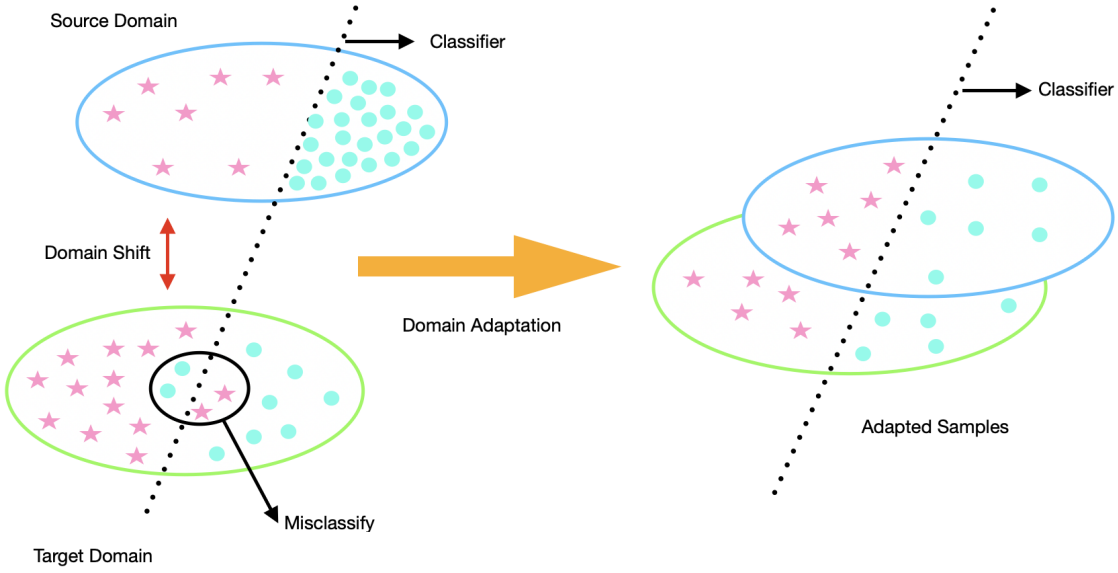

Figure 2.1: Domain shifts across different medical sites (or domains) can impact model performance. This diagram illustrates how the domain shift problem affects a model's performance. On the left, we have the source and target domains. Using a classifier, we can identify misclassified targets in the target domain due to the domain shift problem. However, after applying domain adaptation, we see on the right that there are no misclassified targets between the source and target domains.

In this section, we provide a brief review of domain adaptation techniques in deep learning methods, categorized based on the type of training label datasets: supervised learning, where the entire training dataset is labeled, and semi-supervised learning, where only a portion of the training dataset is labeled. We begin by discussing supervised models, followed by semi-supervised models.

The 3D Md-Unet, a supervised learning model introduced in [109], was designed for collaborative medical image segmentation across multiple datasets, with a focus on bladder cancer segmentation. Based on the Md-Unet backbone [64], the 3D Md-Unet incorporates domain adaptation techniques, including (i) Shared-Specific Adapter (SSA), which used pointwise group convolution for shared feature extraction across datasets, (ii) a shared branch that integrated common features extracted from individual datasets, and (iii) an adaptive weight update strategy to address uncertainty and class imbalance in multi-data collaborative training. These en-

hancements enabled the 3D Md-Unet to extract and share both specific and common features from diverse datasets, including those targeting different organs. The model processed inputs using 3D convolution and a dual-branch structure for feature extraction. The 3D Md-Unet was experimented on the Medical Segmentation Decathlon (MSD) dataset [8].

Similarly, another supervised learning model the 3D U^2-Net was introduced in [76]. It is a universal 3D U-Net that uses domain adaptive techniques for multi-domain medical image segmentation based on the concept of separable convolution, where domain-specific spatial correlations in images are captured through channel-wise convolution, while cross-channel correlations are handled using pointwise convolution. The 3D U^2-Net modified the standard U-Net by incorporating (i) domain adapters which allowed the model to capture domain-specific features while enabling efficient knowledge sharing across multiple domains, and (ii) domain sharing parameters that stored and combined the shared pointwise convolution from individual dataset. During training, the 3D U^2-Net sampled batches from each dataset in a round-robin manner, ensuring that every domain contributed to the shared parameters. The 3D U^2-Net was evaluated on five organ segmentation datasets.

Another, supervised learning model was presented in [186], which introduced a multi-modal learning framework employing domain adaptation techniques for multi-organ segmentation in CT and MRI scans. This approach integrated shared representations within a dual-stream architecture based on a fully convolutional network (FCN) to segment multiple organs across modalities. The shared representation was designed to extract and share common features from multi-modal data, while the dual-stream architecture consists of two components: the first stream that captured modality-specific features (from either CT or MRI), and the second stream that facilitated the exchange of learned information between modalities, enhancing both segmentation accuracy and generalization. The dual-stream structure effectively processed unpaired multi-modal images (where direct correspondences between modalities are absent) while enabling information sharing within a single network. The framework was evaluated on the segmentation of four abdominal organs.

In [52], another multi-modal supervised learning model was introduced. It is an unpaired segmentation via knowledge distillation that enabled information sharing between different modalities. It used modality-specific internal normalization layers to capture information for each modality, which was shared with other modalities by constraining the Kullback–Leibler divergence between the prediction distributions of both modalities. Within the CNN backbone it incorporated two domain adaptation techniques which were (i) Separate Internal Feature Normalization, that normalized input from each modality separately using modality-specific encoders or decoders, with either early or late fusion. It also used a single set of kernels to extract features from all modalities, which improved parameter efficiency and allowed for the extraction of robust, universal representations. And (ii) a Knowledge Distillation Loss, that enhanced knowledge transfer by applying temperature scaling to pre-softmax activations, resulting in softer probability distributions across classes, which facilitates cross learning. The method was implemented in both 2D and 3D architectures and evaluated on two benchmark datasets.

Similarly, another, supervised learning model, DCAC a multi-source domain generalization model for medical image segmentation, based on Domain and Content

Adaptive Convolution was introduced in [75] to tackle segmentation across different imaging modalities. DCAC enhanced the standard U-Net architecture by incorporating domain adaptation techniques, including: (i) a Content Adaptive Convolution (CAC) module, that used a dynamic convolutional head conditioned to capture global image features, enabling the model to adapt to the specific characteristics of the test image; (ii) a Domain Adaptive Convolution, that dynamically adjusted filters based on the input image domain, addressing discrepancies between source domains and unseen target domains; (iii) Feature Aggregation, that integrated feature maps at each layer using Global Average Pooling (GAP); and (iv) a Domain Predictor, that processed multi-scale encoder feature maps, aggregated them using GAP, and concatenated the results into a vector. The DCAC model was evaluated on three datasets across four tasks involving different modalities.

Moving on, the MS-Net, a multi-site supervised learning network designed to enhance prostate segmentation using heterogeneous MRI data, was proposed in [113]. This framework aimed to address limitations associated with single-site samples and variability from different imaging protocols and scanners. Built on a convolutional neural network (CNN) backbone, MS-Net integrated several key components: (i) Domain-Specific Batch Normalization (DSBN), which allows the network to estimate statistics and normalize features separately for each site, effectively addressing inter-site variability; (ii) Auxiliary branches, where data from each site was assigned to an auxiliary branch functioning as an independent feature extractor, enabling the model to learn site-specific knowledge more effectively; (iii) a Universal Network, where, in each iteration, knowledge learned from the auxiliary branches is transferred to the universal network, encouraging shared kernels to capture broader representations, and (iv) Multi-Site-Guided Knowledge Transfer, which enhanced the network's ability to extract generalizable representations from multi-site data, thereby mitigating inter-site differences and improving robustness. The algorithm was evaluated on three heterogeneous prostate MRI datasets from different sites.

DoFE, another supervised learning multi-site, single-modality model leveraging domain adaptation techniques, was introduced in [192]. DoFE is a domain-oriented feature embedding generalizable model using a Knowledge Pool for fundus image segmentation. Within its CNN backbone, DoFE incorporated two domain adaptation techniques: (i) a memory module, inspired by [117] and [166], that dynamically enriched the semantic features of input images by leveraging prior domain knowledge from multiple sources to enhance generalization, and (ii) a domain knowledge pool that stored and retrieved information from various domains during feature embedding, where each entry represents discriminative prototypes from specific training datasets as domain-specific representations. DoFE augmented the input image's features with domain-oriented aggregated features derived from the knowledge pool based on the similarity between the input image and images from multiple source domains. The feature maps were processed through a prediction branch that employed an attention-guided mechanism to dynamically fuse the aggregated features with the original semantic features. The DoFE framework was evaluated on two segmentation tasks involving retinal fundus images from eight sites.

A supervised learning approach incorporating domain-specific batch normalization layers within a convolutional neural network (CNN) backbone was proposed in [94]. This method leveraged shared convolutional filters to facilitate the transfer of learnable features across different domains. Two key domain adaptation techniques

were integrated into the CNN backbone: (i) Using knowledge from prior domains through trained domain-agnostic parameters, enabling rapid fine-tuning of a limited set of domain-specific parameters with minimal risk of overfitting, and (ii) explicitly separating shared and domain-specific parameters, ensuring stable performance on previously learned domains over time. To evaluate the framework, three model variants were developed: (i) Individual Networks, where each domain was trained separately; (ii) a Shared Network, that trained data from all domains using standard batch normalization; and (iii) a Lifelong Multi-Domain Learning Network, that trained data from all domains using domain-specific batch normalization parameters. The framework was tested on four publicly available datasets.

A multi-site supervised learning algorithm for robust white matter hyperintensity segmentation on unseen domains was presented in [218]. The authors aimed to address domain generalization by training the model on samples from various distributions (sources) and testing it on a new, unseen distribution (target) from a different site. The framework used a Mixup techniques that incoporated domain specific adapters per domain to capture and share domain invariant information within a common space. The model was evaluated on the multi-site White Matter Hyperintensity Segmentation Challenge dataset and a private in-house dataset.

A supervised learning method combining domain adaptation (DA) with few-shot learning, referred to as domain adaptation for medical image segmentation, was introduced in [212]. This approach leveraged meta-learning to enable generalization across a diverse range of segmentation tasks. The method integrated domain adaptation and meta-learning within the standard U-Net architecture. The meta-learning component aligned source and target data in a domain-invariant discriminative shared feature space, leveraging the shared knowledge to improve segmentation on domain specific tasks. The algorithm was evaluated on various segmentation tasks from the Medical Segmentation Decathlon [8].

The Multi-Domain Vision Transformer (MDViT), a supervised learning transformer-based backbone was introduced in [53] for small medical image segmentation datasets. The framework employed multi-domain Vision Transformers (ViTs) with domain adapters to reduce data dependency and mitigate Negative Knowledge Transfer (NKT) by adaptively leveraging knowledge from multiple small datasets (domains). Within the ViT backbone, MDViT incorporated a Mutual Knowledge Distillation (MKD) paradigm to enhance representation learning across domains, facilitating knowledge transfer between a universal network and auxiliary domain-specific branches. The MDViT employed two key domain adaptation techniques. First, a Domain Adapter (DA) which integrated domain-specific information into the Multi-Head Self-Attention (MHSA) blocks. The DA employed an attention generation technique to produce domain-aware vectors for each head and an information selection mechanism to adaptively choose the most relevant information for each domain. Secondly, a Mutual Knowledge Distillation (MKD), that enabled bidirectional knowledge transfer between domain-specific networks and the universal network, promoting robust and generalized representations. The domain-specific networks were trained on small, domain-specific datasets, while the universal network learned from all domains. During training, MDViT used a universal network spanning multiple domains with auxiliary branches tailored to each domain. The DA mitigated NKT within MHSA modules, while the MKD strategy facilitates the exchange of domain-specific and shared knowledge, enhancing overall representation learning.

MDViT was experimented on four skin lesion segmentation datasets.

Transitioning to semi-supervised learning the multi-Source domain adaptation for medical image segmentation presented in [147]. The proposed method employed a multi-level adversarial learning strategy to align features across different levels between multiple source domains and the target domain, aiming to improve segmentation accuracy. The model incorporated domain-specific adapters to capture unique information for each domain and domain-shared adapters to share common features among domains. The knowledge gained was used to generate pseudo-labeled images from unseen datasets. The generated data, combined with the original labeled data, were used to train a U-Net-like segmentor, featuring an encoder-decoder structure with skip connections at corresponding levels, in a supervised manner. The framework was evaluated on cardiac and liver segmentation tasks using the Cardiac Dataset [223] and the CHAOS dataset [95].

Similarly, a semi-supervised learning approach called Synergistic Image and Feature Adaptation (SIFA) was introduced in [32] to effectively address the challenge of domain shift. SIFA combined adversarial learning with domain adaptation for cross-modality medical image segmentation. The framework operated in two stages: in the first stage, it incorporated domain adapters within the generator and discriminator to capture domain-specific information, which was shared in a common space across all domains. This process enhanced the domain invariance of the extracted features, which were subsequently used for the segmentation task in the second stage. The framework was evaluated on cross-modality medical image segmentation of cardiac structures.

Another semi-supervised learning approach combining unsupervised domain adaptation (UDA) and zero-shot learning (ZSL) for multi-modality medical image segmentation was presented in [18]. This method leveraged knowledge from a fully annotated image modality to generalize and transfer visual semantics to a new modality with minimal annotation. The framework operated in two stages: in the first stage, a fully supervised model is trained on each modality. In the second stage, a cross-modality adaptation technique was introduced, transferring shared information between modalities using annotated classes. This enabled the zero-shot architecture to inherit relational prototypes from the prior model. The zero-shot model subsequently learned unseen class prototypes and their relationships, enabling segmentation on new modalities. The framework was evaluated on two cross-modality datasets: the CHAOS Challenge [95], and a cardiac dataset from the MMWHS Challenge [224].

Similarly another semi-supervised learning approach, the Prior-aware Neural Network (PaNN) a partially-supervised multi-organ segmentation model, was introduced in [221]. This approach addressed the challenge of background ambiguity in partially labeled datasets by explicitly incorporating anatomical priors related to abdominal organ sizes, leveraging domain-specific knowledge to guide the training process. Within its CNN backbone, PaNN integrated domain-specific knowledge adapters, functioning as auxiliary branches to capture organ-specific information. The model was trained on thirteen CT anatomical structures and evaluated on the MICCAI 2015 Multi-Atlas Labeling Beyond the Cranial Vault challenge dataset [203].

One limitation of domain adaptation (DA) models is the absence of an effective method to prevent the transfer of negative knowledge. Information that is beneficial

in one domain but may adversely impact performance in another, leading to an overall decline in the model's performance.

Reference	Year	Approach	Organ	Modalities	Dimensions	Metrics
[94]	2018	Supervised	Brain	MRI	2D	DS
[186]	2018	Supervised	Multiple	MRI /CT	3D	DS
SIFA [32]	2019	Semi-supervised	Heart	MRI/CT	3D	ASD
PaNN [221]	2019	Semi-supervised	Multiple	CT	2D/3D	DS/HD
3D U^2-Net [76]	2019	Supervised	Multiple	MRI/CT	3D	DS
DoFE [192]	2020	Supervised	Multiple	Fundus	2D	DS/HD
MS-Net [113]	2020	Supervised	Prostate	MRI	2D	DS
[52]	2020	Supervised	Multiple	MRI/CT	2D/3D	DS/HD
[18]	2021	Semi-supervised	Multiple	MRI/CT	3D	DS/ASSD
[212]	2021	Supervised	Multiple	MRI	2D	DS
DANN [218]	2021	Supervised	Multiple	MRI	3D	DS/AVD
DCAC [75]	2022	Supervised	Multiple	Multiple	2D	DS/ASD
3D Md-Unet [109]	2023	Supervised	Multiple	MRI/CT	3D	DS
MDViT [53]	2023	Supervised	Multiple	Dermoscopy	2D	DS/IoU
[147]	2023	Semi-supervised	Multiple	CT	3D	DS/ASD

Table 2.3: A summary of previous work on domain adaptation models, listed in order of year of publication, including the references, year, approach, organ, modalities, image dimensions, and evaluation metrics: Dice Score (DS), average surface distance (ASD), Hausdorff distance (HD), Average symmetric surface distance (ASSD), Absolute volume Difference (AVD), Intersection over Union (IoU), Average Surface Distance (ASD).

2.5 Federated Learning Approaches for Medical Image Segmentation

Federated learning in medical image segmentation allows distributed medical institutions to collaboratively train a shared model while maintaining data privacy. By accessing data from multiple sources, each client can utilize diverse data distributions, addressing the challenge of decentralized data. Information is exchanged between clients in a privacy-preserving manner through an effective interpolation mechanism in a continuous frequency space. The process involves communication between a central server and local clients. In each round, the central server sends global model weights to all clients, who then update the model locally using their data for several epochs. These updated parameters are sent back to the server, which aggregates them to refine the global model, repeating this process until convergence. A visual example of federated learning is provided in Figure 2.2. Several federated approaches have been proposed to tackle the challenge of domain shift in medical image segmentation. Some of the key approaches will be briefly reviewed, categorized by their training methodology, specifically whether or not they used generative adversarial networks (GANs). The review will begin with non-GAN based approaches.

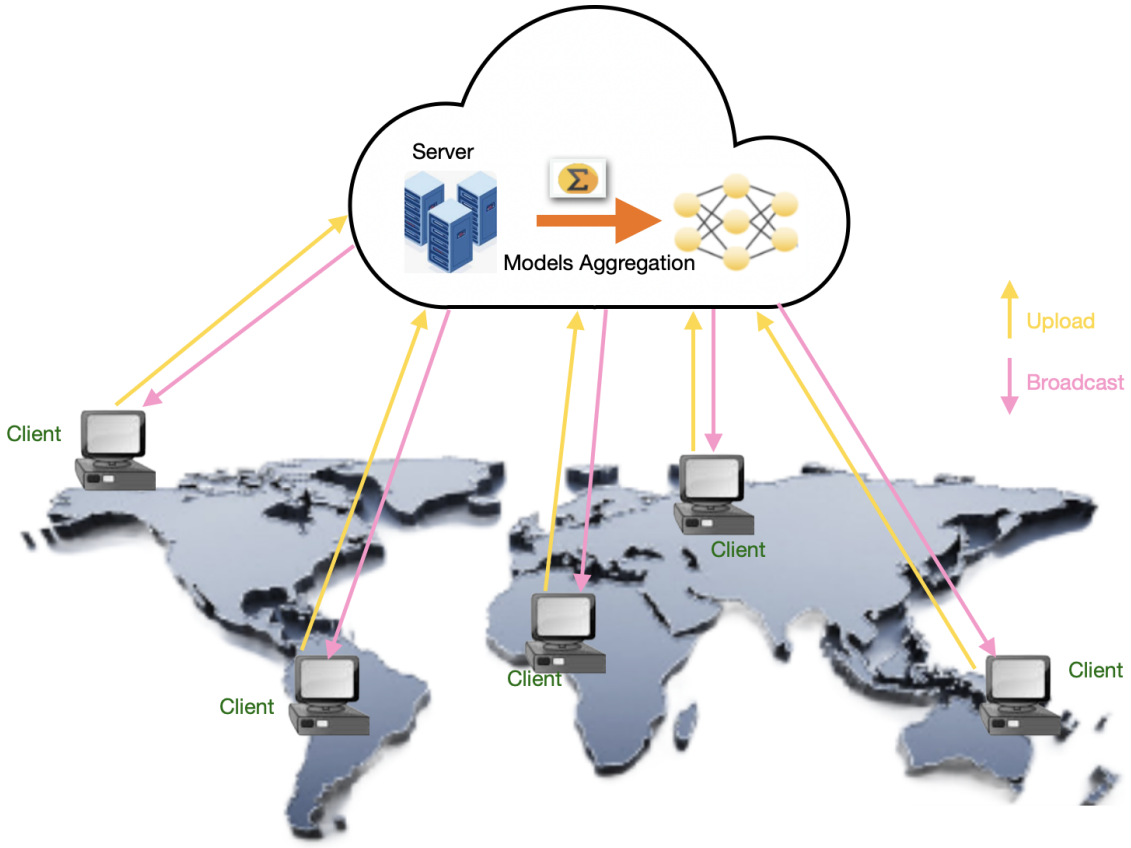

Figure 2.2: An illustration depicting Federated Learning (server-client learning).

Starting with the non-GAN based approaches, a federated domain generalization (FedDG) approach for medical image segmentation using episodic learning in continuous frequency space was introduced in [128]. A framework was presented to train a federated model across multiple distributed source domains, enabling it

to address the challenge of domain shift and generalize effectively to unseen target domains. This approach was built on the widely used Federated Averaging (FedAvg) algorithm, originally developed for natural images and text datasets, which aggregates local model parameters based on the size of each local dataset. The standard FedAvg was enhanced with two key components: (i) continuous frequency space interpolation, where information was shared in the frequency space rather than raw images to preserve privacy, and (ii) boundary-oriented episodic learning, which handled domain distribution shifts by implementing a specialized learning scheme for local training, improving generalization. The method was evaluated on two medical image segmentation tasks: optic disc and cup segmentation on retinal fundus images, and prostate segmentation on T2-weighted MRI scans.

Several researchers have enhanced the Federated Averaging (FedAvg) algorithm to propose non-GAN based approaches for addressing domain shift in medical image segmentation, some of which will be reviewed briefly. The Auto-FedAvg, a learnable federated averaging method, was introduced in [197] to dynamically adjust aggregation weights based on data distributions across silos and the training progress of the models. This approach was applied to COVID-19 lesion segmentation in chest CT and pancreas segmentation in abdominal CT. Similarly, the Whole-brain radiomics for clustered federated personalization in brain tumor segmentation was proposed in [124]. Here the FedAvg was first used to build an initial global model through several communication rounds, and applied within clusters to construct a final model for each cluster of samples with homogeneous texture. The method was evaluated on the FeTS2022 dataset [123]. Moving on, the FedDG, a federated domain generalization approach for medical image segmentation via episodic learning in continuous frequency space, was presented in [112]. FedDG modified FedAvg by transmitting distribution information across clients in a privacy-protecting manner using an effective continuous frequency space interpolation mechanism. The FedDG was evaluated on the BraTS AND Fundus datasets. The abdominal multi-organ segmentation using federated learning, was another approach proposed in [204], combining FedAvg with U-Net for segmenting multiple organs and structures in CT and MRI scans. In this method, each client used a global model based on the U-Net architecture as its local model. To address catastrophic forgetting [59], only the task block of each client's model was fine-tuned while keeping the representation block frozen.

Another non-GAN-based federated learning approach for medical image segmentation, FedSM, was introduced in [201] to address the domain shift problem and reduced the generalization gap between the model and centralized training. A technique called SoftPull was proposed for training federated models. The challenge of domain shift in federated learning was tackled by incorporating an innovative personalized federated learning optimization formulation, which substituted the locally trained model after each training round at the server within the FedAvg [128] framework. The approach was validated on two benchmark datasets.

Similarly, another non-GAN based federated learning approach for distributed medical databases using meta-analysis for large-scale subcortical brain data was introduced in [174] to address the domain shift problem in medical image segmentation. The framework was composed of three main components: (i) a data standardization step, which served as a pre-processing phase to enhance the stability of the analysis and facilitate feature comparison across diverse datasets, (ii) a correction

for confounding factors, employing the Alternating Direction Method of Multipliers (ADMM) to estimate a matrix shared among centers, thereby accounting for confounding variables, and (iii) a variability analysis component, that used Federated Principal Component Analysis (fPCA) to reduce data dimensionality while ensuring privacy by avoiding the sharing of patient information. The framework was evaluated on datasets collected from multiple medical centers.

Moving on, another non-GAN based a feasibility study on brain tumor segmentation, titled multi-institutional deep learning modeling without sharing patient data, was introduced in [168] to address the problem of domain shift. Inspired by federated learning, a U-Net architecture was used as the backbone. The model accepted a single-channel image as input and produced a binary mask of the same size, assigning a class label to each pixel. Two models were proposed: the first, an Institutional Incremental Learning (IIL), which is a collaborative learning approach allowing institutions to train a shared model sequentially. In IIL, each institution trained the model only once using its local methods, requiring minimal bandwidth as the model was transmitted once for training and received twice (once for training and once for the final version). However, IIL faced two significant drawbacks: (a) performance declines as the number of institutions increases, and (b) the risk of catastrophic forgetting [59], where previously learned patterns are lost when new training data replaced old data. The second approach, a Cyclic Institutional Incremental Learning (CIIL), improved upon IIL by cycling through institutions repeatedly and fixing the number of epochs for training at each institution to mitigate catastrophic forgetting. In CIIL, each institution trained the model for a predefined number of epochs before passing it sequentially to the next institution. These algorithms were evaluated on the publicly available BraTS (Brain Tumor Segmentation) dataset [132].

Another non-GAN-based framework, the multi-task federated learning approach for heterogeneous pancreas segmentation, was proposed in [169]. The framework was built on two key innovative features. The first feature, the dynamic task prioritization (DTP), was implemented for multi-task learning by adjusting the weights between different tasks based on an estimation of the key performance index (KPI). Challenging tasks were prioritized by increasing their corresponding weights, while the weights of easier tasks were reduced. The second feature, the dynamic weight averaging (DWA), which served as an optimization approach for multi-task learning tasks [114], focusing on server model aggregation rather than imposing constraints on the loss function. The framework was evaluated on three public datasets.

Transitioning to GAN based approaches, the AsynDGAN, a synthetic learning framework using distributed, asynchronous discriminator generative adversarial networks (GAN) without sharing medical image data, was proposed in [28]. A central generator was trained to learn from distributed discriminators, using the generated synthetic images exclusively to train a segmentation model. By combining GAN with federated learning (FL), a generalizable model was constructed to address the domain shift problem in medical image segmentation across multiple data sources while preserving patient data privacy. The framework included two main components: (i) a central generator, implemented as an encoder-decoder structure for segmentation, incorporating strided convolutions, residual blocks, batch normalization, and activation layers. And (ii) distributed discriminators, trained asynchronously with access only to local data, ensuring data privacy. Each discriminator, deployed at different medical entities (e.g., hospitals), learned to differentiate between real

local images and synthetic images generated by the central generator. The central generator, in turn, captured a joint distribution from the diverse datasets across centers, enabling training for specific segmentation tasks. The AsynDGAN framework was evaluated on multiple datasets.

Similarly, GAN based approach, the Federated Simulation for medical imaging, introduced in [105] to address the domain shift problem, combined Generative Adversarial Networks (GAN) with federated learning for medical image segmentation. It aimed to train a generative model that synthesizes CT volumes and corresponding labels, allowing data sharing without compromising privacy. The framework has two main components, (i) a generative model, that produced an organ shape and material map, independent of imaging devices and (ii) a CT Simulation, which the generated shape and material map are processed through a physics-based CT renderer to create a voxelized label map, generating synthetic CT volumes with labels. During training, the model learned the underlying distribution of the datasets and generates realistic samples, which are then used as auxiliary labeled data for training downstream machine learning models. The Federated Simulation was evaluated on multiple CT datasets from different sites.

Another GAN based federated learning method, termed mixed supervised federated learning for medical image segmentation (FedMix), was introduced in [193] . The FedMix proposed a label-agnostic unified federated learning framework designed for medical image segmentation using mixed image labels to mitigate domain shift problem. In this approach, each client updates the federated model by integrating and effectively utilizing all available labeled data, which ranges from strong pixel-level labels to weak bounding box labels. An adaptive weight assignment procedure was incorporated, allowing each client to learn its aggregation weight during the global model update. FedMix dynamically adjusted each client's aggregation weight, resulting in a rich and discriminative feature representations that increased the diversity and distribution of the model. The FedMix integrated two key features into the U-Net backbone. First the Pseudo Label Generation and Selection (Sample and Refine), which amplified and filtered useful signals from pseudo-supervision by generating pseudo labels through consistency regularization. The pseudo labels were dynamically filtered and refined before being used for training. Secondly, the adaptive aggregation for Federated Model Update (Aggregate), that ensured more reliable clients are assigned higher weights, leading to better model convergence by adjusting client weights based on data quantity and quality. The algorithm was evaluated on two medical image segmentation tasks: breast tumor segmentation (using three public breast ultrasound datasets) and skin tumor segmentation (from four different sources).

Federated learning (FL) has demonstrated success in medical image segmentation, particularly in improving generalization, addressing domain shift, and, most importantly, preserving data privacy. However, many of these methods are based on complex frameworks that combine multiple algorithms, making them challenging to adapt for specific use cases.

Reference	Year	Approach	Organ	Modalities	Dimensions	Metrics
[128]	2017	non-GAN	Brain/Eye	MRI/Fundus	3D/2D	HD
[168]	2019	non-GAN	Brain	MRI	3D	DS
[174]	2019	non-GAN	Brain	MRI	3D	HD
AsynDGAN [28]	2020	GAN	Multiple	MRI/CT	2D	DS/HD
[105],	2020	GAN	Multiple	CT	3D	DS
FedSM [201]	2022	non-GAN	Multiple	CT	2D	DS
FedMix [193]	2022	GAN	Multiple	Multiple	2D	DS
Auto-FedAvg [197]	2021	non-GAN	Chest	CT	2D	Accuracy
[169]	2021	non-GAN	Multiple	CT	3D	DS
FedDG [112]	2021	non-GAN	Brain/Eye	MRI/Fundus	2D/3D	DS
[124]	2024	non-GAN	Brain	MRI	3D	DS
[204]	2024	non-GAN	MRI/CT	Multiple	3D	DS

Table 2.4: A summary of previous work on federated learning models, listed in order of year of publication, including the references, year, approach, organ, modalities, image dimensions, and evaluation metrics: , Hausdorff distance (HD), Dice Score (DS), Accuracy and Intersection over Union (IoU).

2.6 Foundation Model Approaches for Medical Image Segmentation

Foundation models are large-scale, pre-trained models that serve as a base for a wide range of downstream tasks. Trained on massive datasets, they learn general data representations, making them adaptable to different tasks. After the initial training, these models can be fine-tuned using relatively small amounts of task-specific data. Typically, foundation models have billions of parameters and are exposed to extensive datasets, enabling them to capture broad patterns, structures, and representations. Once pre-trained, they can be customized for various applications like segmentation with smaller datasets. These models have shown significant promise in overcoming challenges like generalization and domain shifts between different datasets in medical image segmentation. In this section we will briefly review some foundation models tailored for medical image segmentation categorized base on the backbone: Segment Anything Model (SAM) and non SAM.

The Segment Anything Model (SAM), introduced by researchers at Meta in [97], is a promptable model for image segmentation, trained on a large-scale dataset with 1 billion segmentation masks from 11 million images. It aimed to serve as a foundation model that can generalize across various segmentation tasks through zero-shot transfer. SAM's approach revolves around three key components: task, model, and data. (i) Task: Inspired by zero-shot and few-shot learning in NLP, the authors proposed a promptable segmentation task, allowing SAM to adapt to diverse inputs through prompting. (ii) Model: SAM was built to handle flexible prompts, generate segmentation masks in real-time, and manage ambiguous inputs. (iii) Dataset: The model was trained on a massive dataset of 1 billion segmentation masks, ensuring both privacy and scalability. The model architecture includes: (i) An image Encoder, which is a Vision Transformer (ViT) [51] pre-trained with Masked Autoencoders (MAE) [69] that processed high-resolution images before receiving prompts. (ii) A prompt encoder, that supported sparse prompts (points, boxes, text) using positional encodings and dense prompts (masks) with convolutions, incorporating CLIP's text encoder [151] for text-based inputs. (iii) A mask decoder, that combined image and prompt embeddings using Transformer blocks [27] with self-attention and cross-attention to produce segmentation masks. SAM used a dynamic linear classifier to predict foreground probabilities and was evaluated on 23 unseen datasets. Since its release, numerous SAM variants have been developed, specifically tailored for medical image segmentation, some of which will be reviewed as follows.

The MA-SAM, a modality-agnostic adaptation of the Segment Anything Model (SAM) tailored for 3D medical image segmentation, was introduced in [31]. SAM was initially trained on 2D natural images, which limited its performance in the medical imaging domain due to lack of the third dimension or temporal information. To effectively adapt SAM to medical images, it is crucial to incorporate the third dimensional information or temporal data during fine-tuning. By embedding a series of 3D adapters into the transformer blocks of SAM's image encoder, MA-SAM allowed the model to extract 3D information from the input data. The framework used a parameter efficient fine-tuning strategy that updated only a small portion of weight increments while retaining the majority of SAM's pre-trained weights. To improve SAM's output resolution, which was critical for capturing fine details in medical images, MA-SAM introduced two key enhancements to the decoder: (i) Pro-

gressive upsampling, which added transposed convolutional layers to restore feature maps to the original input resolution and (2) A multi-scale fusion, that used skip connections to combine feature maps from decoder layers with their corresponding encoder layers. Fine-tuning was guided by a hybrid segmentation loss that combines cross-entropy and Dice loss, with training using the Adam optimizer [96] and data augmentation. The The MA-SAM was evaluated on 5 medical segmentation tasks across 11 public datasets (including CT, MRI, and surgical videos).

Similarly, another SAM base model, the SAM-Med2D, a specialized variant of the SAM for 2D medical image segmentation, was introduced in [39]. This adaptation aimed to bridge the gap between natural and medical images by fine-tuning SAM for medical imaging tasks. The SAM-Med2D enhanced SAM's capabilities by incorporating a variety of prompts, including bounding boxes, points, and masks, and fine-tuning both the encoder and decoder. One key modification was the adding of an adapter to SAM's encoder to better capture domain-specific features relevant to medical imaging. SAM-Med2D improved segmentation performance by leveraging SAM's framework, extending its prompt capabilities, and using fine-tuning based on simulated interactive segmentation. The model was trained using simulated interactive segmentation, running through nine iterations per batch. To integrate domain knowledge, the authors curated a dataset with over 4.6 million medical images and 19.7 million masks from both public and private sources and was evaluated on nine medical image datasets

Moving on, the segment anything in medical images (MedSAM) framework, was introduced in [122]. It is a foundation model designed for universal medical image segmentation. It aimed to bridge the gap between general segmentation techniques and the specific needs of medical image segmentation, offering adaptability through user-provided prompts. MedSAM's architecture is similar to the original SAM, featuring an image encoder, a prompt encoder, and a mask decoder. The image encoder mapped input images into a high-dimensional embedding space, while the prompt encoder converts user prompts (e.g., bounding boxes) into feature representations using positional encoding. The mask decoder then combined image and prompt features through cross-attention to generate segmentation results. Built on a Vision Transformer (ViT) backbone, MedSAM used 12 transformer layers with multi-head self-attention and multilayer perceptron (MLP) blocks, all incorporating layer normalization. It was trained on a large-scale medical dataset with 1,570,263 image-mask pairs, covering 10 imaging modalities and over 30 cancer types.

Another SAM based approach, the Localize Anything Model for 3D Medical Images (MedLAM), a one-shot framework for organ and landmark localization in volumetric medical images, was introduced in [102]. It used two self-supervised tasks: Unified Anatomical Mapping (UAM) and Multi-Scale Similarity (MSS). MedLAM was based on the observation that the spatial distribution of organs is consistent across different individuals, and it assumes the existence of a standard anatomical coordinate system in which the same anatomical part in different individuals shares similar coordinates. Thus, it can localize the target anatomy in unannotated scans using similar coordinates. The MedLAM model consisted of three main components: a feature encoder, a multilayer perceptron (MLP), and a feature decoder. The feature encoder and decoder each contained four convolutional blocks. Each encoder block included two convolutional layers followed by downsampling, while each decoder block included two convolutional layers followed by upsampling. The

MLP was composed of three fully connected layers. The model was evaluated on two datasets: (i) the mixed head and neck (HaN) CT StructSeg 2019 dataset [176], which includes 165 volumes from three sources, and (ii) the pancreas CT dataset [159], consisting of 82 volumes.

Similarly, another SAM based approach, the MedLSAM (Localize and Segment Anything Model for 3D CT Images), was introduced in [103], to address the challenge of slice-by-slice annotations typically required by SAM and its medical adaptations, which can be time-consuming as dataset sizes grow. This approach combined MedLAM with SAM to create a 3D localization foundation model capable of identifying any anatomical structure within the body. MedLSAM integrated MedLAM with SAM by using minimal prompts, such as a few annotated extreme points across three directions on template images. This allowed MedLAM to automatically locate the target anatomical region in the entire dataset, generating a 2D bounding box for each image slice. SAM then used the bounding boxes for precise segmentation. The fully automated pipeline involved two stages, first, MedLAM identified the locations of target structures within volumetric medical images, and secondly SAM segmented the structures using the provided bounding boxes. This approach eliminated the need for manual intervention during segmentation. MedLSAM was tested on two 3D datasets covering 38 different organs.

The SAMedOCT, an adaptation of the Segment Anything Model (SAM) specifically designed for 3D segmentation of Retinal Optical Coherence Tomography (OCT) images, was introduced in [56]. The SAMedOCT improved SAM's performance by incorporating the Low-Rank Adaptation (LoRA) technique [103] into the query and value projection layers of each transformer's encoder block. Additionally, it modified the decoder's segmentation head to customize the output for each segmented class, allowing the model to deterministically predict each semantic class and the background, enhancing interpretability and specificity. The fine-tuning process was guided by a combination of cross-entropy and Dice losses, applied to downsampled ground since SAMedOCT's output has lower spatial resolution. The training process used the AdamW optimizer [96], with a warm-up period and exponential learning rate decay. SAMedOCT was evaluated on the MICCAI 2017 RETOUCH challenge dataset [21], consisting of 112 OCT volumes.

The self-sampling meta SAM (SSM-SAM), was introduced in [104], to enhance few-shot medical image segmentation through meta-learning for rapid online adaptation. It was built on SAM's feature extraction capabilities, leveraging its zero-shot potential without requiring extensive training data. SSM-SAM used MAML++ [9], a meta-learning method originally trained on natural text and images, as its foundation. The framework has three main modules which are (i) An online fast gradient descent optimizer, that was enhanced with a meta-learner for quick adaptation to new tasks. (ii) A self-sampling module, that generated well-aligned visual prompts to improve attention focus. (iii) An attention-based decoder, designed to capture relationships between slices for effective few-shot segmentation. The SSM-SAM's structure was divided into two parts, the first part was a SS-SAM, which is a modified version of SAM that replaced its prompt encoder and mask decoder with a self-sampling prompt encoder and a Flexible Mask Attention Decoder (FMAD). It also included adapters in the image encoder to enhance learning from new tasks, improving transferability. The second part, was the SSM-SAM, which is built on SS-SAM by adding a meta-learning-based optimizer, further boosting

SAM's performance in few-shot segmentation tasks. During training, the SAM image encoder remains frozen, with a learnable adapter added to each transformer layer for parameter-efficient learning. A self-sampling operation refined image embeddings before passing them to the FMAD, that produced the predicted mask. The MAML++ meta-learner optimized the initial parameters for quick adaptation to different organs. The framework was evaluated on abdominal CT and MRI datasets, and experimental.

The medical SAM Adapter (Med-SA), introduced in [196], adapted the Segment Anything Model (SAM) for medical image segmentation by incorporating domain-specific knowledge using a lightweight adaptation approach. Instead of fine-tuning all of SAM's parameters, Med-SA retained most of SAM's pre-trained weights and introduced targeted adaptations. Med-SA's key innovations include: (i) A space-depth transpose (SD-Trans), that adapted SAM for 3D medical images by transposing spatial dimensions into the depth dimension, allowing SAM's self-attention blocks to process 2D and 3D data. SD-Trans has two branches: the space branch to capture spatial correlations and the depth branch to capture depth correlations, integrating depth information into the spatial attention output. (ii) A hyper-prompting adapter (HyP-Adpt), which is a prompt-conditioned adaptation that used visual prompts to generate weight maps, enabling richer interactions between prompts and model embeddings. During training, the Med-SA used click prompts (positive for foreground and negative for background) and bounding boxes (BBox). It began with random click sampling and used an iterative sampling method to simulate real user interactions, placing new clicks in error regions to refine segmentation. Med-SA was evaluated on 17 medical image segmentation tasks across various modalities, including CT, MRI, ultrasound, fundus, and dermoscopic images.

The SAMAug another SAM based approach was introduced in [215], to enhance medical image segmentation by using SAM-generated data for input augmentation. While SAM's initial segmentation for medical images may not be high-quality, the masks, features, and stability scores it produces can improve other segmentation models. SAMAug integrated these elements into a three-step process: First was the segmentation and boundary prior maps, for which SAM-generated masks were used to create boundary prior maps. Each training image is augmented by adding the segmentation prior map as a second channel and the boundary map as a third channel, forming an enriched training set. The second step was a model trained with SAM-augmented images, the augmented images were used to train medical segmentation models like U-Net, allowing them to benefit from both the raw image data and the additional information provided by SAM's masks. The final step was the model deployment with SAM-augmented images. During testing, models trained with SAM-augmented data processed test inputs similarly. If a model was trained on both raw and augmented data, the outputs from both were averaged to improve accuracy. The SAMAug was evaluated on three datasets for different segmentation tasks.

Another SAM based method is the ESP-MedSAM framework, an efficient self-prompting SAM designed for universal domain-generalized in medical image segmentation, was introduced in [202]. This method addressed the high computational costs, reliance on manual prompts, and challenges in generalization in clinical scenarios. The ESP-MedSAM, included several innovations to enhance performance and generalization. First, it developed a Multi-Modal Decoupled Knowledge Distil-

lation (MMDKD) strategy to construct a lightweight, semi-parameter-sharing image encoder that generated discriminative visual features for multiple modalities. Next, it introduced the Self-Patch Prompt Generator (SPPG), which automatically generated high-quality dense prompt embeddings to guide segmentation. Finally, it designed the Query-Decoupled Modality Decoder (QDMD), that provided an independent decoding channel for each modality to prevent conflicts during the segmentation process. The ESP-MedSAM was evaluated on six medical imaging modalities.

Other methods that have employed SAM for medical image segmentation includes: An Extensive studies using datasets from various imaging modalities to compare the performance of SAM variants was presented in [167]. This approach employed the two point-prompt strategies which are: (i) A multiple positive prompts, where one prompt was placed near the centroid of the target structure, while others were randomly positioned within the structure, and (ii) a combined positive and negative prompts, where one positive prompt was placed near the centroid of the target structure and two negative prompts were positioned outside it, maximizing the distance from the positive prompt and each other. These strategies were evaluated on 24 unique organ-modality combinations across 11 publicly available MRI, CT, ultrasound, dermoscopy, and endoscopy datasets. Additionally, the interactive 3D medical image segmentation with SAM 2, combining zero-shot and SAM, was proposed in [170] and evaluated on the BraTS2020 and MSD datasets. Another, SAM-based image enhancement (SAM-IE), was introduced in [188], which leveraged SAM-generated masks and features to enhance image quality for disease diagnosis, by combining binary and contour masks generated by SAM. This method was tested on four medical image datasets. Moving on, an experimental study was presented in [125], that generated point and box prompts for SAM using a standard method simulating interactive segmentation. This study was evaluated on 19 medical imaging datasets, and concluded that SAM demonstrated impressive zero-shot performance on some datasets but moderate results on others. Next, a fine-tuning approach for SAM using few examples, termed cheap lunch for medical image segmentation, was proposed in [57]. The approach incorporated an exemplar-guided synthesis module and Low-Rank Adaptation (LoRA) fine-tuning strategy. The algorithm was evaluated on the MRI brain tumor (BraTS) and multi-organ CT segmentation (Synapse) datasets. Moving on, the AdaptiveSAM was introduced in [144]. This approach incorporated bias-tuning that required less than 2% of SAM's trainable parameters and negligible expert intervention, using free-form text prompts to segment objects of interest. It was evaluated on three datasets from diverse modalities, including ultrasound and X-ray. Finally, a comprehensive overview of recent efforts to extend SAM's efficacy to medical image segmentation tasks, encompassing empirical benchmarking and methodological adaptations, was presented in [214].

Transitioning to non-SAM approaches, the One-Prompt Model [195] was introduced for universal medical image segmentation, combining the strengths of one-shot and interactive models to address real-world clinical needs. The architecture consisted of a standard CNN-based image encoder and a sequence of One-Prompt Formers for segmentation tasks. Three inputs were utilized: a query image, a template image, and a prompt for the template, with the model predicting the segmentation of the query image. Skip connections were employed to integrate multi-scale features in both the encoder and decoder stages. Key components of the architecture included: (i) an encoder that processed query and template images through the same U-Net

encoder to extract features, (ii) a One-Prompt Former that combined downsampled query features with prompt embeddings and template features using attention blocks, (iii) a cross-attention mechanisms where the query branch integrated skip-connected query embeddings while the template branch used a Prompt-Parser to merge prompts with template features, (iv) a final cross-attention and Feedforward Neural Network (FNN) that unified the two branches to transfer the template's segmentation to the query, and (v) a Prompt-Parser that generated an adaptive attention mask to focus on the prompted target. The model was trained using 64 open-source medical datasets and over 3,000 clinician-labeled prompts.

Large foundation models in medical image segmentation have demonstrated significant advancements in segmentation tasks. However, they often require substantial computational resources, which are typically inaccessible to many researchers. A notable limitation of these approaches is the lack of strategies to dynamically reduce image resolution during training, which could effectively decrease space complexity and improve computational efficiency. Additionally, most of these models are trained on natural images or limited medical image datasets, leading to poor generalization on downstream medical imaging tasks. Developing large foundation models trained exclusively on diverse medical image datasets would be a significant step forward in addressing these challenges.

Reference	Year	Backbone	Organ	Modalities	Dimensions	Metrics
MedLAM [102]	2021	SAM	Multiple	CT	3D	IoU
[57]	2023	SAM	Multiple	MRI/CT	2D	DS/HD
SAM-Med2D [39]	2023	SAM	Multiple	Multiple	2D	Accuracy
MedLSAM [103]	2023	SAM	Multiple	CT	3D	DS/HD
SAMedOCT [56]	2023	SAM	Eye	OCT	3D	DS/AVD
Med-SA [196]	2023	SAM	Multiple	Multiple	2D	DS/IoU
SAMAug [215]	2023	SAM	Multiple	Multiple	2D	DS
[144]	2024	SAM	Multiple	X-Ray/US	2D	DS/IoU
One-Prompt [195]	2024	ViT/CNN	Multiple	Multiple	2D	DS
MA-SAM [31]	2024	SAM	Multiple	MRI/CT	3D	DS/HD
SSM-SAM [104]	2024	SAM	MRI/CT	Multiple	2D	DS
MedSAM [122]	2024	SAM	Multiple	Multiple	2D/3D	DS
ESP-MedSAM [202]	2024	SAM	Multiple	Multiple	2D	DS/HD
[214]	2024	SAM	Multiple	Multiple	2D	DS
[167]	2024	SAM	Multiple	Multiple	2D	DS
[170]	2024	SAM	Multiple	MRI/CT	2D	DS/HD
[188]	2024	SAM	Multiple	Multiple	2D	AUC
[125]	2024	SAM	Multiple	Multiple	2D	IoU

Table 2.5: A summary of previous work on foundation models, listed in order of year of publication, including the references, year, method, organ, image dimensions, and evaluation metrics: Intersection over Union (IoU), Dice Score (DS), Hausdorff distance (HD), Accuracy, Absolute volume Difference (AVD), Area under the curve (AUC).

2.7 Fine-Tuning Approaches for Medical Image Segmentation

Fine-tuning is a common strategy to improve the generalisation of deep learning models, particularly when adapting a pre-trained model to specific tasks or datasets in medical image segmentation. In this section, we will review several fine-tuning approaches used to enhance medical image segmentation, organized by their architectural backbone: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and a hybrid combinations of both.

Annotated datasets from diverse sources often exhibit class overlaps due to inconsistent label definitions, such as differing annotations for the same organ across datasets. For instance, the heart's annotation in one dataset may overlap with the aorta in another. To address this challenge in medical image segmentation, a CNN-based algorithm, MultiTalent, was proposed in [184], integrating three key modifications into the nnU-Net framework. A class-adaptive loss function, combining Binary Cross-Entropy loss with Dice loss, was employed to manage overlapping classes while preserving label properties. A one-hot label vector was introduced to handle label contradictions by enabling independent class prediction. Additionally, a sigmoid activation function was used to allow the prediction of multiple classes per pixel, accommodating overlapping class representations. During training, the model used a shared backbone with independent segmentation heads for each class, preserving the unique characteristics of each dataset's labels. MultiTalent supported two primary scenarios: (i) Combined Multi-Dataset (MD) Training, that developed a foundational model capable of segmenting all classes from partially annotated datasets, and (ii) Pre-Training, that fine-tuned the foundational model's learned representations for new tasks. The framework was trained on 13 public abdominal CT datasets comprising 1,477 3D images and 47 classes.

Similarly, another CNN-based algorithm, the Med3D, was introduced in [36] as a transfer learning approach for 3D medical image analysis, aimed at developing pre-trained models using diverse medical image datasets that could be fine-tuned for downstream tasks. A heterogeneous Med3D network was designed to extract general 3D features across varied medical domains. The Med3D focused on two main objectives, the first was to build a comprehensive 3D medical dataset called 3DSeg-8 by aggregating smaller datasets from multiple medical domains. And the second, was train baseline networks (pre-trained models) on this dataset for subsequent transfer and fine-tuning to other tasks. The network was based on the ResNet architecture [68], a variant of the U-Net with skip connections. An encoder-decoder structure was employed, with the encoder connected to eight separate decoder branches, each optimized for a specific dataset within 3DSeg-8. During training, only the decoder branch relevant to the active dataset was used, while others remained inactive. For testing, the decoder was removed, and the trained encoder was applied for transfer learning to new tasks. Data augmentation techniques were used to balance smaller datasets to match the size of the largest dataset. The pre-trained Med3D models were evaluated on tasks such as lung segmentation, pulmonary nodule classification, and liver segmentation.

Another CNN-based backbone, Models Genesis, was introduced in [222] as a framework for training generic models for 3D medical imaging using a self-supervised learning approach that does not require labeled data. Comprehensive image repre-

sentations were learned through a combination of self-supervised tasks, consolidated into a single image restoration task using an encoder-decoder architecture, making the framework scalable and well-suited for transfer learning and fine-tuning across various 3D medical imaging tasks. The primary goal was to develop transferable image representations capable of generalizing across different diseases, organs, and imaging modalities. The self-supervised training process involved cropping sub-volumes from patient CT images, applying transformations to these sub-volumes, and training the model to restore the transformed sub-volumes to their original form. Using a 3D U-Net backbone, the framework incorporated four key components: (i) an image restoration module that mapped transformed sub-volumes back to their original state using an encoder-decoder network with skip connections, (ii) a non-linear transformation that preserved the appearance of anatomical structures with a monotonic intensity transformation function, (iii) a local pixel shuffling to introduce texture variations and enhance boundary detection, and (iv) an outer and inner cutouts that promoted contextual understanding by masking specific regions while exposing others. Models Genesis was pre-trained on 623 chest CT scans from the LUNA 2016 dataset [86].

Moving on, another CNN-based fine-tuning model, the Interactive Medical Image Segmentation framework using deep learning with image-specific fine-tuning, was proposed in [190]. This interactive framework was designed with a bounding box and image-specific fine-tuning-based CNN segmentation network. Several key innovations were incorporated into the framework as follows: (i) a bounding box and scribble-based binary segmentation within the CNN backbone to extract the region of interest (ROI), (ii) an image-specific fine-tuning to adapt the CNN model to each test image independently, (iii) a weighted loss function that accounted for network and interaction-based uncertainty during the image-specific fine-tuning process, and (iv) a zero-shot learning technique for segmentation employed within the CNN backbone. A single model was trained, which was fine-tuned for different downstream tasks by first using bounding boxes to extract the ROI and then applying zero-shot learning for segmentation within the CNN backbone. The framework was evaluated on several brain MRI datasets.

The 3D Anisotropic Hybrid Network (AH-Net), another CNN based approach, was introduced in [115] as a fine-tuning transfer learning method in which convolutional features learned from 2D images were transferred to 3D anisotropic volumes. Strong generalization capabilities from features learned on B-scan slices were leveraged, and inter-slice information was effectively used through focal loss [158]. The 2D fully convolutional ResNet [148] (a U-Net variant with skip connections) was extended into a 3D architecture by adding an extra dimension to the 2D kernel, and skip connections were incorporated between the feature encoder and decoder. To enable multi-scale feature extraction, a pyramid volumetric pooling module [217] was integrated at the end of the decoder path, just before the classification layer. AH-Net was evaluated on two datasets: a private Digital Breast Tomosynthesis dataset and the public Liver Tumor Segmentation Challenge dataset [19].

The use of convolutional neural networks (CNNs) for medical image analysis with fully fine-tuning was introduced in [179]. This approach was designed to analyze how the availability of training samples influenced the decision between using pre-trained CNNs or training CNNs from scratch. It was demonstrated that fine-tuning a pre-trained CNN in a layer-wise manner resulted in incremental performance im-

provements. The approach was validated on four medical image datasets.

The fine-tuning of U-Net for ultrasound image segmentation was proposed in [5]. In this approach, a large model was trained using both natural and medical images and subsequently fine-tuned for medical image-specific segmentation tasks. The standard U-Net architecture was modified by replacing the transposed convolutional layers with bilinear upsampling followed by 2×2 convolution. The algorithm was trained on a large dataset comprising natural images, breast ultrasound scans, and chest X-rays. Data augmentation techniques were applied to increase the size of the ultrasound and X-ray datasets to match the scale of the natural image dataset.

Transitioning to Vision Transformer-based architectures, the DAFT framework, a data-aware fine-tuning approach for foundation models aimed at efficient and effective medical image segmentation, was proposed in [150]. Based on the demographic characteristics of the input image, meta-learning was utilized to efficiently select the most suitable pre-trained model for fine-tuning from a pool of 11 Vision Transformer (ViT) pre-trained models. The DAFT framework was evaluated on multiple MRI and CT medical image datasets.

Similarly, another Transformer based approach for fine-tuning was presented in [211]. It is a deep stacked transformations model that can be fine-tuned for downstream task-specific problems in medical image segmentation. The framework applied a series of transformations to each image during network training to simulate domain shifts within specific medical imaging modalities. A pre-trained model was trained on an augmented large dataset, termed BigAug. The pre-trained model generalized effectively to unseen domains. The framework comprised of two key components. The first was deep stacked transformations, where a sequence of image transformations was applied, with each transformation characterized by a probability of application and a magnitude function. The second component was a 3D deep segmentation backbone based on Vision Transformers (ViT), which transferred deep features learned from large-scale 2D images to a 3D encoder-decoder network. During training, transformations were applied to each mini-batch to simulate domain-specific shifts, generating augmented data with corresponding annotations that altered image quality, appearance, and spatial configurations. Sub-volumes were randomly cropped from the whole volume and segmented into masks with one-channel annotations. The sub-volumes were evenly distributed between the foreground and background to enhance data diversity. At inference, a sliding window with overlap was used across the entire 3D volume to produce the final segmentation. The algorithm was evaluated on four publicly available 3D prostate MRI datasets, three 3D heart MRI datasets, and one 3D ultrasound dataset

Another Vision Transformer-based approach, LiteMedSAM, a low-rank adaptation and multi-box efficient inference method for medical image segmentation, was introduced in [116]. LiteMedSAM adjusted the probabilities of selecting each modality during data loading to address severe imbalances in modality data, improving segmentation performance for medical images with limited data in certain modalities. Within the Vision Transformer backbone, LiteMedSAM incorporated a low-rank adaptation technique into the multi-head attention and multilayer perceptron components to fine-tune the model for downstream medical image segmentation tasks. The approach was evaluated on a large dataset encompassing multiple modalities, including MRI, CT, PET, Ultrasound, X-Ray, Dermotology, Endoscopy, Fundus, and Microscopy.

Transitioning to hybrid-based approaches, UniSeg, a method combining CNN and Vision Transformer (ViT), was proposed in [207] as a prompt-driven universal segmentation model and strong representation learner. This approach was designed to fine-tune task-specific downstream medical image segmentation tasks, leveraging the one-hot label vectors to create a single model capable of segmenting medical images across diverse sources and modalities. Built for the segmentation of multiple organs, tumors, and vertebrae in 3D medical images across various modalities and domains, UniSeg introduced a learnable universal prompt to capture correlations among all tasks. This universal prompt, combined with image features, was converted into a task-specific prompt and fed into the decoder, allowing the model to become task-aware early and enhancing task-specific training in the decoder. Based on the nnU-Net backbone, UniSeg is comprised of a vision encoder, a fusion and selection (FUSE) module, and a prompt-driven decoder. During training, the FUSE module generated task-specific prompts, enabling the model to adapt to the ongoing task. The universal prompt and vision encoder features were passed to the FUSE module, which selected the appropriate task-specific prompt based on the current task. The task-specific prompt was later introduced in the decoder to improve task-specific processing, allowing a single decoder and segmentation head to predict various targets under their corresponding ground truths. The algorithm was pre-trained on 11 upstream datasets containing 3,237 volumetric scans from three modalities (CT, MR, and PET) with targets covering eight organs and fine-tuned on two downstream datasets.

Similarly, another hybrid approach, Hermes, a context-prior learning framework aimed at universal medical image segmentation, inspired by radiology residency programs, was proposed in [60]. The framework sought to develop foundational models for medical image segmentation that could be fine-tuned to downstream tasks by leveraging the diversity and commonalities across various clinical targets, body regions, and imaging modalities. Hermes addressed the challenges related to data heterogeneity and annotation inconsistencies through a universal approach that performed multiple tasks within a single model. The framework consisted of five main components: (i) an oracle-guided context-prior learning that explicitly learned context-prior knowledge alongside the segmentation backbone using diverse medical imaging datasets, (ii) a task context prior that treated each task as a binary segmentation task, enabling flexibility in handling diverse datasets, incomplete annotations, and conflicting class definitions, (iii) a modality context prior that adapted to the unique characteristics of medical images sourced from multiple modalities, (iv) a conditioned segmentation that used a prior fusion module to merge context-prior tokens with image features through attention mechanisms within CNN and Transformer architectures, employing bi-directional cross-attention, and (v) the hierarchical modeling that enhanced segmentation performance by integrating multi-scale contextual prior knowledge from posterior prototypes at various scales. Hermes was evaluated on 2,438 3D images from eleven diverse datasets across five modalities (CT, PET, T1, T2, and cine MRI) covering multiple body regions for both standard and pre-trained fine-tuning models.

Similar to foundation models, fine-tuning medical image segmentation models has shown significant progress in advancing segmentation tasks. However, it often demands substantial computational resources that are typically inaccessible to many researchers. A notable limitation of these approaches is the absence of strategies

to dynamically reduce the image resolution during training, which could effectively lower space complexity and enhance computational efficiency.

Reference	Year	Approach	Organ	Modalities	Dimensions	Metrics
[179]	2016	CNN	Multiple	Multiple	2D	AUC
[190]	2018	CNN	Brain	MRI	2D	DS
AH-Net [115]	2018	CNN	Multiple	X-Ray/CT	3D	DS
Med3D [36]	2019	CNN	Multiple	MRI/CT	3D	DS
[5]	2020	CNN	Multiple	X-Ray	2D	DS
BigAug [211]	2020	ViT	Multiple	MRI/US	2D/3D	DS
Models Genesis [222]	2021	CNN	CT	Multiple	3D	AUC
MultiTalent [184]	2023	CNN	Multiple	CT	3D	DS/HD
UniSeg [207]	2023	CNN/ViT	Multiple	CT/MR/PET	3D	DS
Hermes [60]	2024	CNN/ViT	Multiple	CT/MR/PET	3D	DS
DAFT [150]	2024	ViT	Multiple	MRI/CT	3D	DS
LiteMedSAM [116]	2024	ViT	Multiple	Multiple	2D/3D	DS

Table 2.6: A summary of previous work on fine-tuning models in medical image segmentation, listed in order of year of publication, including the references, year, method, organ, image dimensions, and evaluation metrics: Area under the curve (AUC), Dice Score (DS), Hausdorff distance (HD), Accuracy and Intersection over Union (IoU).

2.8 Publicly Available Multi-Source Datasets for Medical Image Segmentation

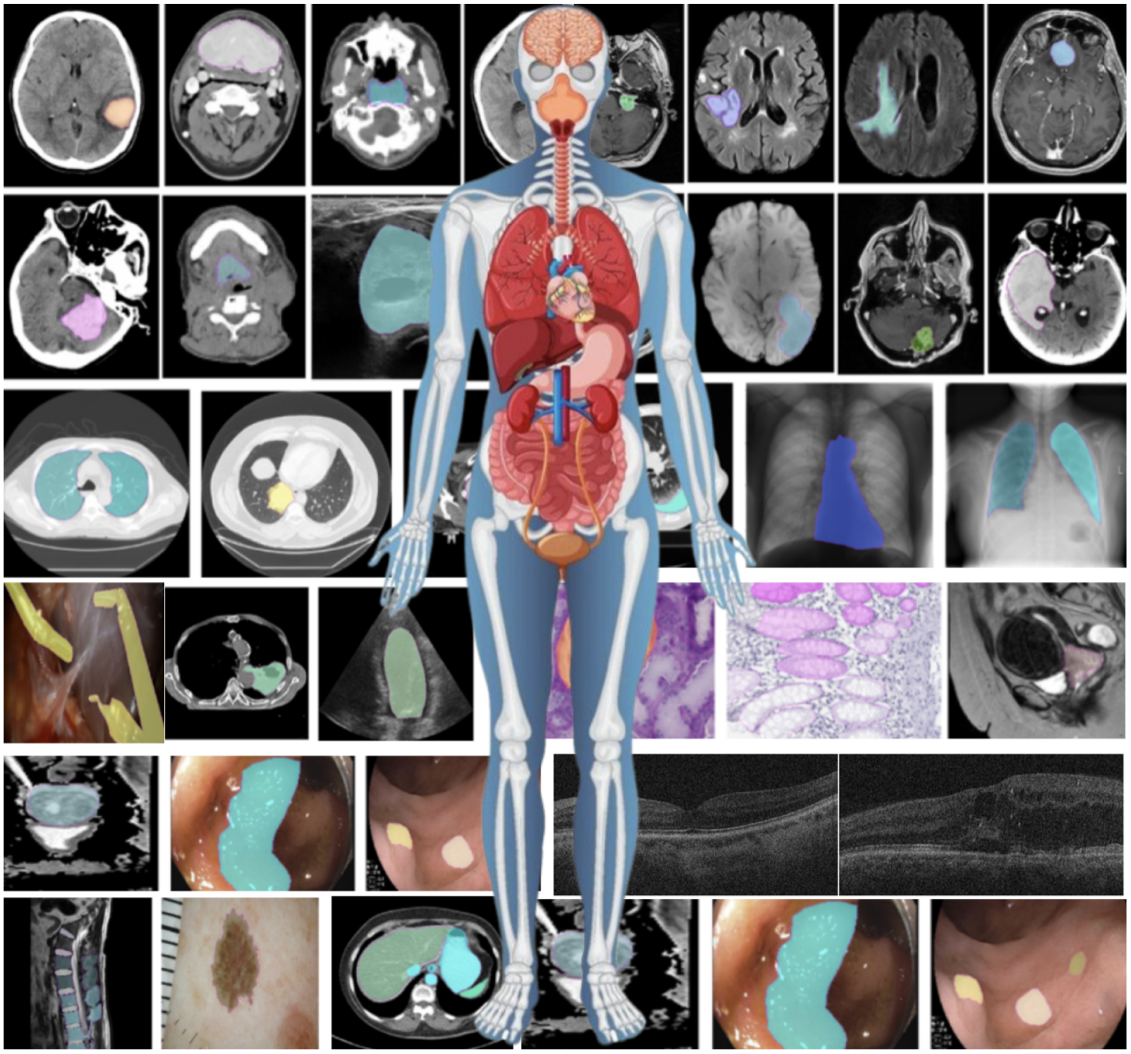

Figure 2.3: An illustration showcasing the significant variability across datasets from various organs and sources.

One of the main challenges in deep learning for medical imaging is the limited availability of large annotated datasets. This is due to strict privacy regulations surrounding medical data, as well as the labor-intensive and time-consuming nature of annotating such datasets. However, numerous small annotated datasets from diverse sources spanning various imaging modalities, organs, and disease types are publicly available. An illustration showcasing the significant variability across datasets from various organs and sources is shown in Figure 2.3. Some researchers have created larger public datasets by combining data from these sources, and some of these will be briefly discussed in this section.

The 3DSeg-8 dataset [36] is composed of eight public medical datasets, covering various organs and tissues, with either CT or MR scans. To expand the dataset, three data augmentation techniques were applied: translation, rotation, and scaling.

The Multi-Modality Whole Heart Segmentation (MM-WHS) challenge dataset, introduced in [225], consists of 120 multi-modality whole heart images collected from various clinical sites. The dataset includes 60 cardiac CT and 60 cardiac MRI scans, all captured in real clinical environments. The images encompass the entire heart, extending from the upper abdomen to the aortic arch. The cardiac CT data were acquired using two CT scanners (Philips Medical Systems, Netherlands) following a standard coronary CT angiography protocol at two sites in Shanghai, China. The cardiac MRI scans were obtained from two hospitals in London, UK: St. Thomas Hospital using a 1.5T Philips scanner (Philips Healthcare, Best, The Netherlands) and Royal Brompton Hospital using a Siemens Magnetom Avanto 1.5T scanner (Siemens Medical Systems, Erlangen, Germany).

The whole-body FDG-PET/CT dataset with manually annotated tumor lesions, introduced in [63], consists of 1,014 publicly available studies collected between 2014 and 2018 as part of a prospective registry study. These studies involve 900 patients, with 501 cases featuring malignant lymphoma, melanoma, or non-small cell lung cancer (NSCLC), and 513 studies serving as negative controls, without PET-positive malignant lesions. The dataset was acquired using a Biograph mCT PET/CT scanner (Siemens, Healthcare GmbH, Erlangen, Germany) at the University Hospital Tübingen in Germany, and annotations were provided by a radiologist and nuclear medicine specialist in a clinical setting.

The Head and Neck Tumor Segmentation (HECKTOR) dataset, described in [7], is a multi-center resource comprising of 883 3D PET and CT volumes. These images were collected from 9 medical centers across 4 countries (Canada, Switzerland, France, and the USA) using 16 different imaging devices. The dataset includes annotations for three segments: background (value 0), primary Gross Tumor Volumes (GTVp) (value 1), and nodal Gross Tumor Volumes (GTVn) (value 2), with lymph nodes grouped under the same label. Additionally, the dataset contains patient information such as age, gender, weight, tobacco and alcohol consumption, performance status, HPV status, and details about treatments, including radiotherapy, chemotherapy, and/or surgery.

The CTSpine1K dataset, introduced in [111], is curated from four open-source datasets, comprising a total of 1,005 CT volumes (over 500,000 labeled slices and more than 11,000 vertebrae) with diverse appearance variations. It includes 33 anatomical categories spanning 7 partially-labeled datasets with approximately 2,800 volumes. These 33 anatomies cover 3 pelvic bones, 5 abdominal organs, and 25 vertebrae. The four datasets are:
1) COLONOG [92]: Consists of 825 CT scans collected from 15 medical sites worldwide between February 2005 and December 2006.
2) HNSCC-3DCT-RT [15]: A subset of 31 head-and-neck squamous cell carcinoma (HNSCC) patients with 3D CT scans taken pre-, mid-, and post-treatment, acquired using a Siemens 16-slice CT scanner under standard clinical protocols. 3)Medical Segmentation Decathlon (MSD) [8]: Contains ten distinct datasets.
4) COVID-19: Includes 40 chest CT scans from 632 COVID-19 patients. The images were acquired during an outbreak, with RT-PCR confirmation for SARS-CoV-2 infections.

The AMOS dataset, a large-scale abdominal multi-organ benchmark for versatile medical image segmentation, is introduced in [89]. It comprises 500 CT and 100 MRI scans collected from multiple centers, vendors, modalities, and phases, featuring patients diagnosed with abdominal tumors or abnormalities at Longgang District People's Hospital and Longgang District. Each scan includes voxel-level annotations for 15 abdominal organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus.

SA-Med2D-20M, introduced in [206], is a large-scale dataset compiled from 140 public and private datasets across 10 medical centers. It includes 10 imaging modalities, 4 anatomical structures and lesion types, and covers 31 major human organs. In total, the dataset comprises 4.6 million 2D medical images and 19.7 million corresponding masks with 219 labels, spanning nearly the entire body and demonstrating extensive diversity. 3D volumes were sliced into 2D images, further expanding the dataset size.

The Medical Segmentation Decathlon (MSD) dataset [8] is a publicly available collection comprising ten distinct datasets, with a total of 2,633 images representing various anatomies and modalities from institutions worldwide. Each image is annotated with one to three regions of interest (ROIs), covering a total of 17 target areas. The datasets are individually annotated and include the following:
1) Brain: This dataset includes 750 multiparametric MRI scans from patients with glioblastoma or lower-grade glioma. The imaging sequences are native T1-weighted (T1), post-Gadolinium T1-weighted (T1-Gd), native T2-weighted (T2), and T2 Fluid-Attenuated Inversion Recovery (FLAIR). The target ROIs are the tumor subregions: edema, enhancing tumor, and non-enhancing tumor. The data overlaps with the 2016 and 2017 Brain Tumor Segmentation (BraTS) challenges [132], [13], [14].
2) Heart: This dataset comprises 30 mono-modal MRI scans of the entire heart taken during a single cardiac phase, using free breathing with respiratory and ECG gating. The target ROI is the left atrium. This data was part of the 2013 Left Atrial Segmentation Challenge (LASC) [181].
3) Hippocampus: This dataset contains 195 MRI images from 90 healthy adults and 105 adults with non-affective psychotic disorders. The images are T1-weighted MPRAGE scans, with the target ROIs being the anterior and posterior hippocampus, including the hippocampus proper and parts of the subiculum. The data was acquired at Vanderbilt University Medical Center, Nashville, US.
4) Liver: This dataset consists of 201 contrast-enhanced CT images from patients with primary liver cancers and metastatic liver disease due to colorectal, breast, or lung cancers. The target ROIs include the liver and tumors within it. The data was obtained at IRCAD Hôpitaux Universitaires, Strasbourg, France, and includes a subset from the 2017 Liver Tumor Segmentation (LiTS) challenge [19].
5) Lung: This dataset includes preoperative thin-section CT scans from 96 patients with non-small cell lung cancer. The target ROI is the tumors within the lung. The data was sourced from the Cancer Imaging Archive `https://www.cancerimagingarchive.net/`.
6) Prostate: This dataset comprises 48 prostate multiparametric MRI (mp-MRI) studies, including T2-weighted, Diffusion-weighted, and T1-weighted contrast-enhanced series. The target ROIs are the prostate peripheral zone (PZ) and transition zone

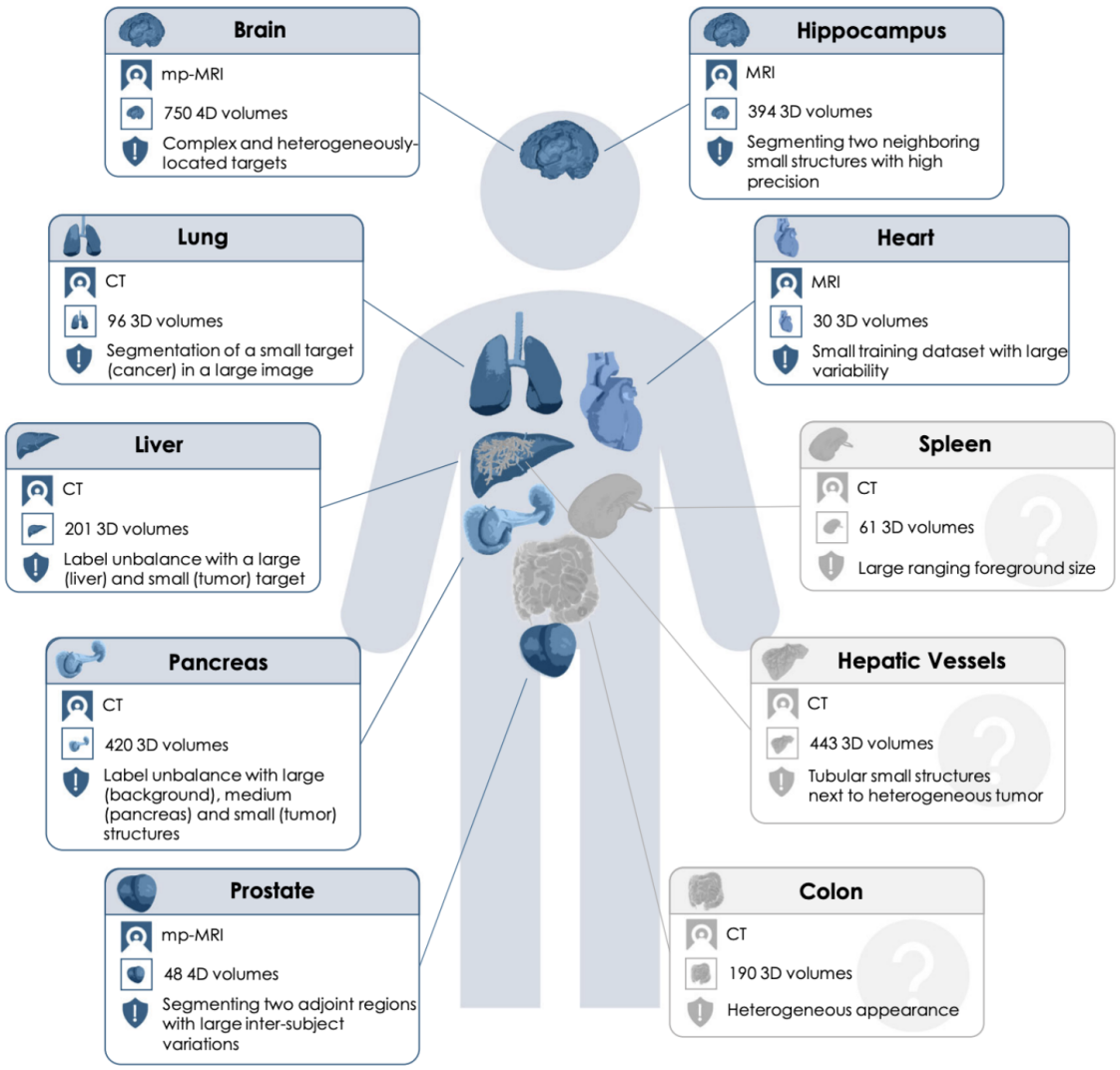

Figure 2.4: The Medical Segmentation Decathlon (MSD) dataset [8] provides a comprehensive collection of different target regions, imaging modalities, and challenging characteristics. It is divided into seven known tasks (in blue, representing the development phase: brain, heart, hippocampus, liver, lung, pancreas, prostate) and three mystery tasks (in gray, representing the mystery phase: colon, hepatic vessels, spleen). The dataset includes MRI (magnetic resonance imaging), mp-MRI (multiparametric MRI), and CT (computed tomography) scans. The image is sourced from [8].

(TZ). The data was acquired at Radboud University Medical Center, Nijmegen, Netherlands.
7) Pancreas: This dataset contains 420 portal-venous phase CT scans of patients undergoing pancreatic mass resections. The target ROIs include the pancreatic parenchyma and pancreatic masses (cysts or tumors). The data was collected at Memorial Sloan Kettering Cancer Center, New York, US. 8) Colon: This dataset includes 190 portal-venous phase CT scans of patients undergoing resection of pri-

mary colon cancer. The target ROI is the primary colon cancer lesions. The data was acquired at Memorial Sloan Kettering Cancer Center, New York, US.
9) Hepatic Vessels: This dataset consists of 443 portal-venous phase CT scans from patients with various primary and metastatic liver tumors. The target ROIs are the hepatic vessels and tumors within the liver. The data was sourced from Memorial Sloan Kettering Cancer Center, New York, US.
10) Spleen: This dataset includes 61 portal-venous phase CT scans from patients undergoing chemotherapy for liver metastases. The target ROI is the spleen. The data was acquired at Memorial Sloan Kettering Cancer Center, New York, US.
A summary of the Medical Segmentation Decathlon (MSD) dataset is shown in figure 2.4.

Reference	Year	Organ	Modality	Dimension	Device	Centre
3DSeg-8 [36]	2019	Multiple	Multiple	3D	Multiple	Multiple
MM-WHS [225]	2019	Heart	CT/MRI	3D	Multiple	Multiple
HECKTOR [7]	2021	Head/Neck	PET/CT	3D	Multiple	Multiple
FDG-PET/CT [63]	2022	Whole body	PET/CT	3D	Single	Single
CTSpine1K [111]	2022	Multiple	CT	3D	Multiple	Multiple
AMOS [89]	2022	Multiple	CT/MRI	3D	Multiple	Multiple
MSD [8]	2022	Multiple	Multiple	3D	Multiple	Multiple
SA-Med2D-20M [206]	2023	Multiple	Multiple	2D	Multiple	Multiple

Table 2.7: A summary table of publicly available large datasets from diverse sources, listed in order of publication year. It includes information such as references, year, target organ, imaging modality, image dimensions, number of vendor devices, and the number of data collection centers.

Reference	Code
PaNN [221]	https://github.com/DLTK/DLTK
3D U^2-Net [76]	https://github.com/huangmozhilv/u2net_torch/
Med3D [36]	https://github.com/Tencent/MedicalNet
AsynDGAN [28]	https://github.com/tommy-qichang/AsynDGAN
PIPO-FAN [55]	https://github.com/DIAL-RPI/PIPO-FAN
[52]	https://github.com/carrenD/ummkd
[105]	https://nv-tlabs.github.io/fed-sim/
MS-Net [113]	https://github.com/liuquande/MS-Net
TransUNet [34]	https://github.com/Beckschen/TransUNet
MedLAM [102]	https://github.com/LWHYC/RPR-Loc
SegFormer3D [149]	https://github.com/OSUPCVLab/SegFormer3D
DANN [218]	https://github.com/xingchenzhao/MixDANN
DoDNet [209]	https://git.io/DoDNet
Omni-Seg [44]	https://github.com/ddrrnn123/Omni-Seg
Swin UNETR[66]	https://monai.io/research/swin-unetr
CoTr[199]	https://github.com/YtongXie/CoTr
FedDG [112]	https://github.com/liuquande/FedDG-ELCFS
FedMix [193]	https://github.com/Jwicaksana/FedMix
UNETR [67]	https://github.com/Project-MONAI/research-contributions
MedFormer [61]	https://github.com/yhygao/CBIM-Medical-Image-Segmentation
DCAC [75]	https://github.com/ShishuaiHu/DCAC/
CLIP-Driven [110]	https://github.com/ljwztc/CLIP-Driven-Universal-Model
MDViT [53]	https://github.com/siyi-wind/MDViT
MultiTalent [184]	https://github.com/MIC-DKFZ/MultiTalent
UniSeg [207]	https://github.com/yeerwen/UniSeg
[144]	https://github.com/JayParanjape/biastuning
Hermes [60]	https://github.com/yhygao/universal-medical-image-segmentation
[17]	https://github.com/BiQiWHU/DFQ
One-Prompt [195]	https://github.com/MedicineToken/one-prompt
Models Gen. [222]	https://github.com/MrGiovanni/ModelsGenesis
FedSM [201]	https://github.com/NVIDIA/NVFlare/examples/FedSM
MA-SAM [31]	https://github.com/cchen-cc/MA-SAM
MedLSAM [103]	https://github.com/openmedlab/MedLSAM
Med-SA [196]	https://github.com/MedicineToken/Medical-SAM-Adapter
SAMAug [215]	https://github.com/yizhezhang2000/SAMAug
[124]	https://github.com/MatthisManthe/radiomics_CFFL
SSM-SAM [104]	https://github.com/DragonDescentZerotsu/SSM-SAM
MedSAM [122]	https://github.com/bowang-lab/MedSAM
ESP-MedSAM [202]	https://github.com/xq141839/ESP-MedSAM
[214]	https://github.com/YichiZhang98/SAM4MIS
[170]	https://github.com/Chuyun-Shen/SAM_2_Medical_3D

Table 2.8: A summary table of authors who have made their code publicly available, along with their corresponding GitHub links, is provided.

2.9 Summary

In this section, we provided a brief overview of key approaches reflecting recent advancements in medical image segmentation from diverse data sources for medical image analysis. These approaches are categorized into specific designed models, universal model approaches, domain adaptation techniques, federated learning, foundation models, and fine-tuning methods. The reviewed literature highlights significant progress in disease diagnosis, segmentation, and generalization using deep learning methods, with notable commonalities across different domains. However, there remain gaps within each category, which will be summarized as follows:
1) **Specific designed models**: This section provided brief overview of models tailored or designed for a specific task. The approaches were categorized based on the architectural backbone employed: CNN, U-Net, ViT, or a combination of these. One notable limitation of this approach their task-specific nature. They will perform well on a particular task, organ, or disease type but may struggle or perform poorly when applied to other tasks or problem.
2) **Universal Model**: This section provided a brief overview of recent key approaches aimed at enhancing model generalizability by increasing the diversity of training data through the integration of data from multiple diverse sources to develop a single universal model. The approaches were categorized based on the backbone architecture: convolutional neural network (CNN), vision transformer (ViT), or a hybrid combination of both. Notable limitations of this approach include the size of the training datasets and the imbalance in dataset sizes and modalities. As demonstrated in [51], both CNNs and ViTs require a substantial amount of training data. Additionally, since most datasets are sourced online and collected from various medical centers, there is significant variability in their sizes, resulting in imbalances in both dataset size and modalities.
3) **Domain Adaptation**: This section provided a brief overview of key medical image segmentation techniques that have incorporated domain adaptation (DA) strategies to address the domain shift problem when training on data from multiple sources and applied the model to unseen data. The approaches were categorized based on the training label datasets: supervised learning and semi-supervised learning. One notable limitation of this approach is the lack of an effective mechanism to prevent the transfer of negative knowledge. Information that is beneficial in one domain but detrimental in another, which can ultimately lead to a decline in the model's overall performance.
4) **Federated learning**: This section provided an overview of key approaches addressing domain shift in federated learning for medical image segmentation, enabling distributed medical institutions to collaboratively train a shared model while preserving data privacy. The approaches were categorized based on the training methodology: generative adversarial networks (GANs) and non-GANs based. One notable limitation of this approach is that many of the methods are based on complex frameworks that combine multiple algorithms, making them challenging to adapt for specific use cases.
5) **Foundation models**: This section provides a brief overview of key foundation models, which are large-scale, pre-trained models that serve as a base for a wide range of downstream tasks in medical image segmentation categorized base on the backbone: Segment Anything Model (SAM) and non SAM. These models are

trained on massive datasets, enabling them to learn general data representations and adapt to various tasks. However, this approach has several notable limitations: (i) Foundation models often require substantial computational resources, which are typically inaccessible to many researchers, (ii) many approaches lack strategies to dynamically reduce image resolution during training, which could decrease space complexity and improve computational efficiency, and (iii) most models are trained on natural images or limited medical image datasets, resulting in poor generalization to downstream medical imaging tasks.
6) **Fine-tuning models**: Fine-tuning is a widely used strategy to improve the generalization of deep learning models, particularly when adapting a pre-trained model to specific tasks or datasets in medical image segmentation. This section provides a brief review of key fine-tuning approaches used to enhance medical image segmentation, organized by their architectural backbone: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid combinations of both. Similar to foundation models, fine-tuning approaches for medical image segmentation have notable limitations, including: (i) they often require substantial computational resources, which are typically inaccessible to many researchers, and (ii) the lack of strategies to dynamically reduce image resolution during training, which could lower space complexity and improve computational efficiency. A table summarising the approaches, gaps and limitations is provided in Table 2.9 provides a summary of the approaches, gaps, and limitations.

Also, we have provided a summary of large benchmark medical image datasets from diverse sources and links of GitHub repositories of researchers who have made their code public.

In the next three chapters, we will explore potential solutions to enhance the detection, segmentation, and generalizability of diseases in medical images across diverse data sources. Chapter 3 introduces a novel algorithm, nnUNet_RASPP, which integrates an Atrous Spatial Pyramid Pooling (ASPP) block to capture global contextual information and residual connections to address overfitting into the nnU-Net framework for retinal disease segmentation. While Chapter 4, extends this concept by incorporating into the baseline backbone three key innovations leading to the development of a novel algorithm: Deep_ResUNet++. Finally, Chapter 5, introduces two novel algorithms: (i) MMIS-Net (MultiModal Medical Image Segmentation Network): a transfer learning approach designed for medical image segmentation across diverse data sources, modalities, organs, and disease types, and (ii) CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-Specific Batch Normalization): a hybrid combination of Convolutional Neural Networks for feature extraction, and Vision Transformers to capture long-range dependencies, while incorporating domain-specific adapters to extract domain specific information and address the challenge of negative knowledge transfer.

Approaches	Gaps	Limitations
Specific Designed Models	Models tailored for specific tasks, categorized based on architectural backbone: CNN, U-Net, ViT, or combinations.	Task-specific nature models perform well for particular tasks, organs, or diseases but may struggle with other applications.
Universal Model	Enhances generalizability by integrating diverse training data from multiple sources. Categorized based on backbone: CNN, ViT, or hybrid.	Imbalances in dataset sizes and modalities due to sourcing from multiple medical centers [51]. It also, requires substantial training data,
Domain Adaptation	Incorporates domain adaptation (DA) strategies to mitigate domain shift issues in medical image segmentation. Categorized based on training labels: supervised and semi-supervised learning.	Lack of an effective mechanism to prevent negative knowledge transfer, which can degrade the model's performance.
Federated Learning	Addresses domain shift in federated learning, allowing distributed institutions to collaboratively train models while preserving privacy. Categorized based on methodology: GAN-based and non-GAN-based.	Complex frameworks combining multiple algorithms, making adaptation for specific use cases challenging.
Foundation Models	Large-scale, pre-trained models serving as a base for diverse downstream tasks. Categorized based on backbone: SAM-based and non-SAM-based models.	(i) Requires substantial computational resources. (ii) Lacks dynamic resolution reduction strategies for training efficiency. (iii) Often trained on natural images or limited medical datasets, leading to poor generalization.
Fine-tuning Models	Fine-tuning improves model generalization by adapting pre-trained models to specific medical segmentation tasks. Categorized based on backbone: CNN, ViT, and hybrid models.	(i) Requires substantial computational resources. (ii) Lacks strategies for dynamic resolution reduction to improve space complexity and efficiency.

Table 2.9: A comparative table summarizing the approaches, gaps, and limitations discussed in this review.

Chapter 3

Enhancing Retinal Disease Detection, Segmentation, and Generalization with an ASPP Block and Residual Connections Across Diverse Data Sources

Deep learning methods have been successful in the detection and diagnosis of diseases in medical images. However, most of these methods are trained and tested on data from the same sources, resulting in poor generalization performance when applied to new, unseen data sources, as is often required in real-world scenarios. One major cause of the lack of generalization is the high variability in the quality of the images, stemming from diverse sources collected using different manufacturer devices or scanners, following varying protocols, and by experts with varying levels of expertise. One way to circumvent this challenge is to build a single, generalizable model by combining data from multiple sources. In this chapter, we modified the nnU-Net [84] architecture by integrating an Atrous Spatial Pyramid Pooling (ASPP) block [35] to : (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. Adding residual connections to address overfitting, proposing an enhanced variant termed nnUNet_RASPP (nnU-Net with Residual and Atrous Spatial Pyramid Pooling). Additionally, we conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed. Assessing the advantages of nnUNet_RASPP over the original nn-UNet, the algorithms were validated on the MICCAI 2017 RETOUCH challenge dataset, which includes data from three device vendors across three medical centers, focusing on patients with two types of retinal diseases. Experimental results on the hidden test set show that the proposed nnUNet_RASPP outperformed the baseline nnU-Net, current state-of-the-art algorithms, and large foundation models for medical imaging by a significant margin. It achieved a mean Dice Score (DS) of 0.823 across the three retinal fluids, with scores of 0.84, 0.80, and 0.83 for intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelium detachment (PED), respectively. Additionally, we achieved a perfect Area Under the

Curve (AUC) score of 1 for detecting the presence of fluid in all three fluid classes. Furthermore, we currently hold the top rank on the MICCAI 2017 RETOUCH challenge online leaderboard: `https://retouch.grand-challenge.org/Results/`, with the best overall performance for both online and offline evaluations.

The work presented in this chapter is published in [138], and [136]. These manuscripts are lead-authored by the author of this thesis, who made substantial contributions to the conception, data collection, processing, and writing, as well as sole contributions to the implementation and result analysis.

3.1 Introduction

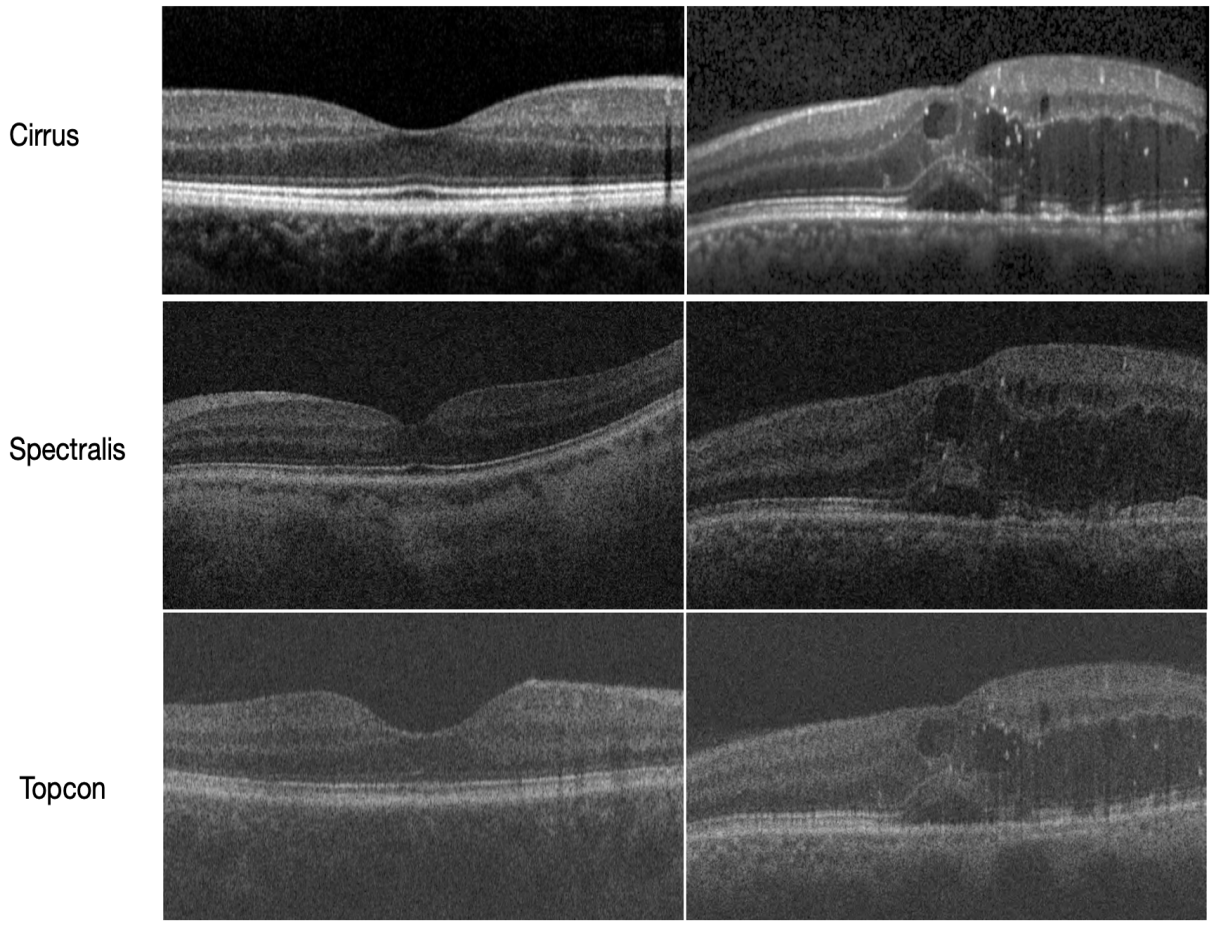

Figure 3.1: An illustration of typical image variability across three manufacturers: Cirrus, Spectralis, and Topcon, showing both a normal retina and a retina with macular edema.

A recent study [106] indicates a rise in retinal diseases, including Age-related Macular Degeneration (AMD) and Diabetic Macular Edema (DME), in Europe,

with over 34 million people affected by AMD and 4 million by DME. AMD is most common among individuals aged 50 and above, with early stages being asymptomatic and progressing slowly to more severe stages. DME, characterized by retinal thickening due to intraretinal fluid accumulation in the macula, is prevalent among diabetic patients. Retinal Optical Coherence Tomography (OCT), a noninvasive imaging technique, provides cross-sectional scans of the eye with qualitative 3D visualization of the retinal anatomy, aiding in the study of retinal structure and the detection of diseases. OCT is the primary imaging tool for retinal analysis and detecting diseases due to its high 3D quality. However, OCT images often suffer from motion artifacts, which lower the signal-to-noise ratio (SNR) due to speckle noise. Also, there is a trade-off between SNR and spatial resolution thus some manufacturers acquire multiple B-scans at the same anatomical location to reduce noise on the expense of producing fewer B-scans. To address this issue, device manufacturers must balance high SNR, image resolution, and scanning time, resulting in varying image quality across different vendors. An illustration demonstrating the high variability in image quality is shown in Figure 3.1.

To develop a high-performance automated model that generalizes well across images from various devices, we employ the nnU-Net framework [84] and an enhanced version called nnUNet_RASPP. Our main contributions are as follows:
1) We enhanced the nnU-Net architecture by integrating an Atrous Spatial Pyramid Pooling (ASPP) block to: (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. Also, we added residual blocks to address overfitting, thereby improving the model's generalization performance. These enhancements address the challenges of high image variability associated with diverse, multi-source datasets, such as that of the Retouch challenge. As a result, nnUNet_RASPP demonstrates improved robustness and a greater ability to generalize across datasets with varying anatomical structures and image quality.
2) We conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed.

The remainder of this chapter is organized as follows: Section 3.2 introduces the proposed nnUNet_RASPP and highlights its differences from the standard nnU-Net. Section 3.3 provides a brief overview of the leading methods in the RETOUCH Grand Challenge. The dataset, experimental results, comparisons, and visualizations are presented in Section 3.4. Finally, the summary is presented in Section 3.5.

3.2 Methods

In this section, we introduce the nnU-Net and our proposed enhanced variant, nnUNet_RASPP. We explain the modifications made, how they differ from the standard nnU-Net, and how these changes improve performance.

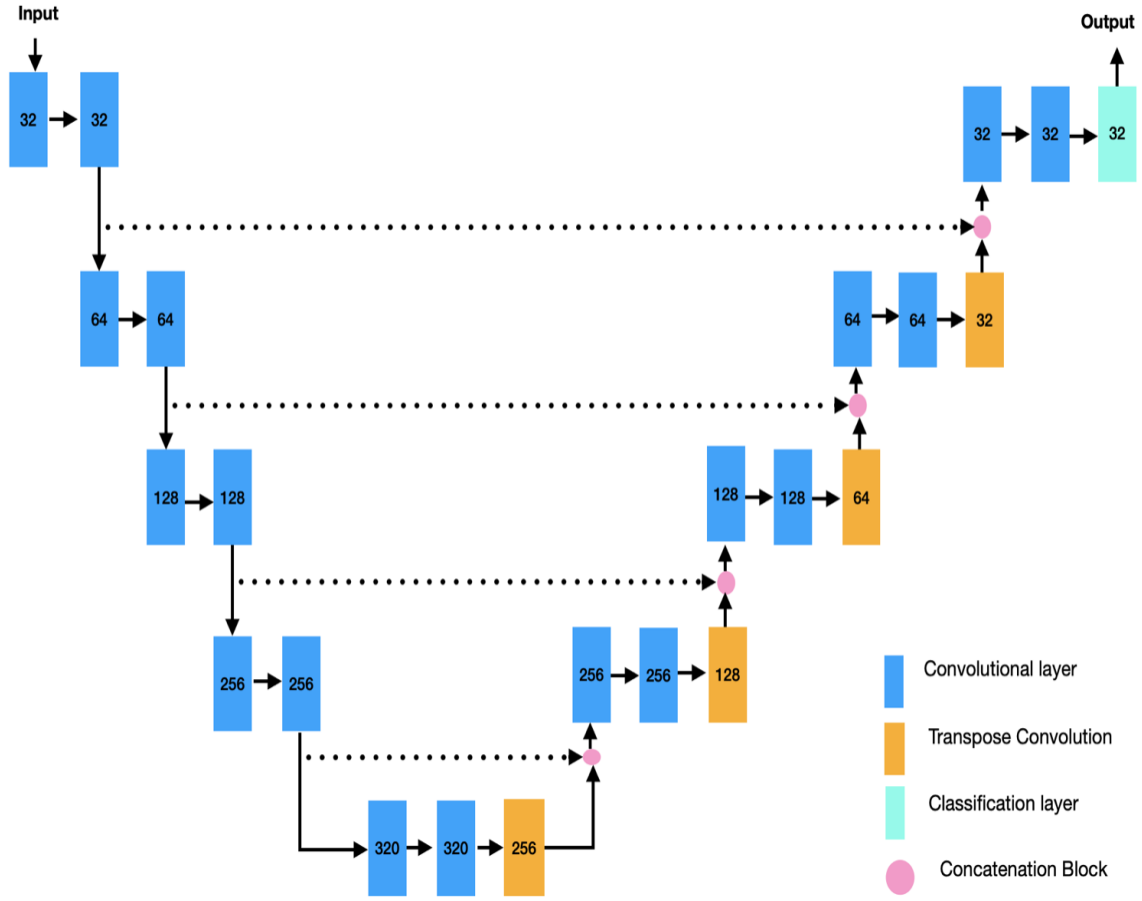

Figure 3.2: An illustration of the standard U-Net architecture used in nnU-Net.

3.2.1 U-Net

The U-Net is an end to end architecture for medical image segmentation. It consists of 3 main parts: the encoder, the decoder and bottleneck between the encoder and decoder. The encoder captures contextual information (or features extraction) and reduces the size of the feature map by half after every convolutional block as we move down the encoding path by implying strided convolutions. Pixels localisation is done at the decoder through up-sampling. As we move up the decoder path the size of the feature map is doubled after every convolutional block by implying transposed convolutions, and for the reconstruction process features maps are concatenated to the corresponding map in the encoder path using up-sampling operations. The bottleneck serves as a bridge, linking the encoding and decoding paths together. It consists of a convolutional block that ensures a smooth transition from the encoder path to the decoder path. At the encoding path, decoding path and bridge layer each convolutional block consists of a convolutional layer that converts the pixels of the receptive field into a single value before passing it to the next operation followed

by an instance normalisation to prevent over-fitting during training. This is followed by a LeakyReLU activation function to diminish vanishing gradient. A high level diagram to illustrate the architectural structure of the standard U-Net is shown in Figure 3.2.

3.2.2 nnU-Net and nnUNet_RASPP

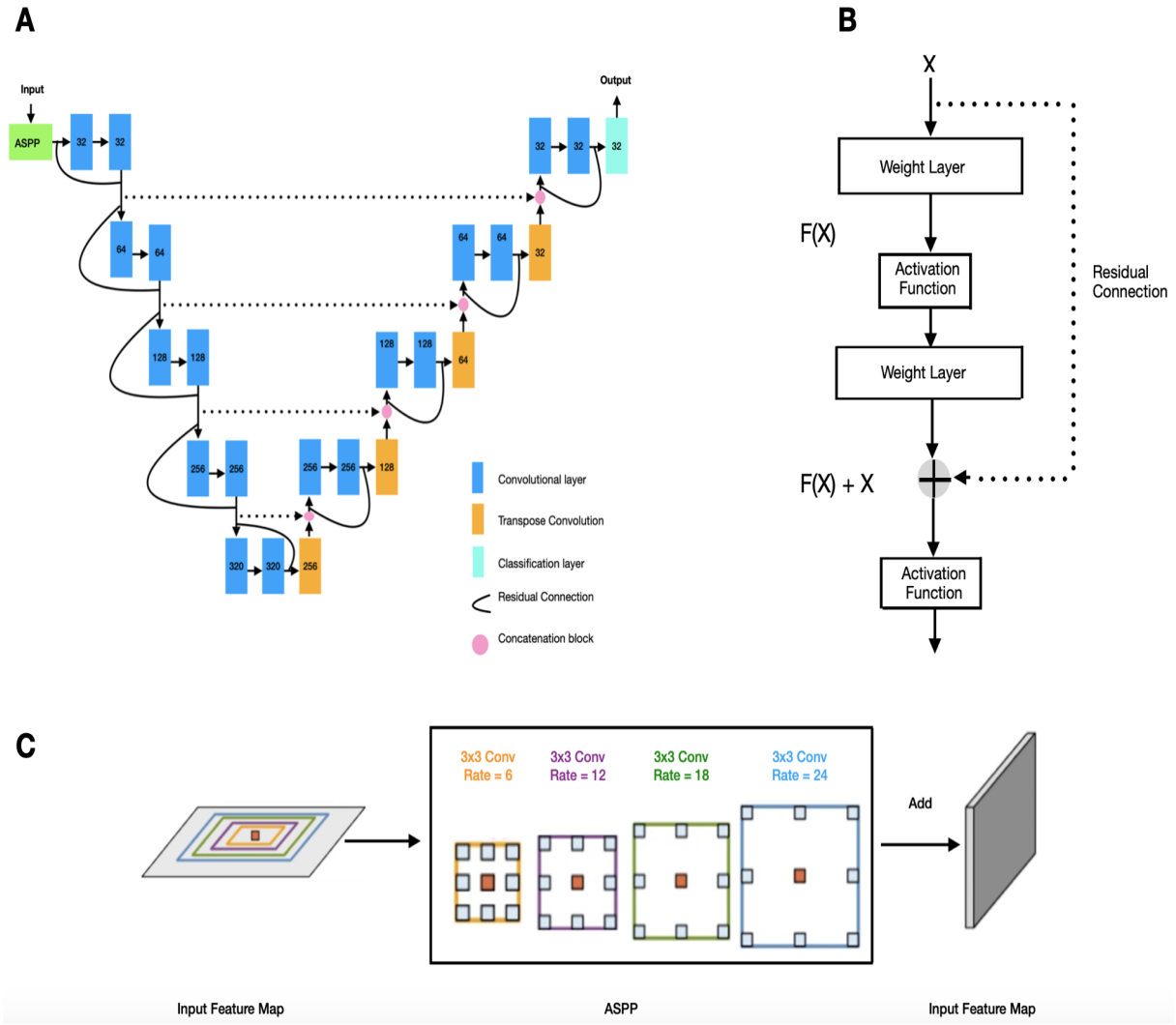

Figure 3.3: A high level illustration of nnUNet_RASPP architecture with B, a residual connection block [68] to address the vanishing gradient problem where X is an input and F(X) is a function of X, and C, an ASPP block [35] of multiple parallel filters at different dilating rates or frequencies to capture global information.

The nnU-Net [84] is a self-configuring and automatic pipeline for medical image segmentation with the ability to automatically determine and choose the best model hyper-parameters given the data and the hardware availability, thus alleviating the problem of trial and error of manual parameters setting. Given a training data the framework extracts the data-fingerprint such as modality, shape, and spacing and base on the hardware (GPU memory) constraints the network topology, image resampling methods, and input-image patch sizes are determined. After training is

complete, the framework determines if post-processing is needed. The framework uses the standard U-Net as the network's architecture.

Inspired by the success of nnU-Net [84] we have introduced an enhanced architecture nnUNet_RASPP by incorporating an ASPP block [35] and residual connections in the network's architecture to solve the problem of data source variation.

1) **ASPP**: It is a technique used to extract or capture global contextual features by applying parallel filters with different dilation rates to a given input filter. Incorporating, ASPP enables the nnUNet_RASPP to: (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. These improvements address the challenges of high image variability associated with diverse datasets such a the Retouch dataset, thereby enhancing nnUNet_RASPP's overall robustness and ability to generalize across diverse datasets with varying anatomical structures and image quality. Within the nnUNet_RASPP, the ASPP block was incorporated at the input layer of the encoding path. Given the high variability of fluid classes, where the three fluid types exhibit no fixed locations or shapes, placing the ASPP block at the input layer is essential. This strategic positioning enables the model to capture global contextual features early in the process, before downsampling, thereby enhancing its ability to generalize across diverse and unpredictable fluid patterns. The diagram of the ASPP block is demonstrated in Figure 3.3.C.

2) **Residual Connections**: It is a technique used to address the problem of vanishing gradients. The U-Net architecture employs the chain rule for backpropagation during training, which can sometimes result in vanishing gradients. One way to mitigate this issue is by introducing residual connections into the network's architecture. These connections help reduce the training error rate as the network's depth increases. The nnU-Net automatically determines the optimal depth of the network, and incorporating residual connections further decreases the training error rate, enabling the network to learn complex features and enhancing overall performance. Residual connections were integrated into every convolutional layer along both the encoding and decoding paths to lower the training error rate, facilitate the learning of complex features, and combat the problems of vanishing gradients and overfitting. The diagram of residual connection is demonstrated in Figure 3.3.B.

Incorporating these techniques into the standard nnUNet improves the overall performance of the network. The diagram of nnUNet_RASPP is demonstrated in Figure 3.3.A.

3.3 RETOUCH Grand Challenge Overview

The RETOUCH grand challenge [21] is a competition focused on the segmentation and detection of three retinal fluids from retinal OCT images. The training dataset comprises of 70 raw images with corresponding masks, while the testing dataset includes 40 raw images without their corresponding masks. To ensure fairness in comparison, the organizers employed a blinded evaluation by retaining the masks or ground truth of the testing dataset, and participants can submit their predictions via email for evaluation. In adherence to competition requirements, each submission must be accompanied by a written paper explaining the methods employed. The results of the submission are communicated to the teams via email and are also published on the organizer's website alongside the accompanying papers. The RETOUCH challenge, initially organized in conjunction with MICCAI 2017 in Quebec, Canada, featured the participation of eight teams. Subsequently, the competition transitioned to an online format, and it remains ongoing, continuing to accept submissions [156].
In this section we will provide a brief overview of other methods that are in the leading positions from the RETOUCH competition.

3.3.1 SAMedOCT

The SAMedOCT [56] is inspired and adpated from the Segment Anything Model (SAM) [97]. It is a foundation model for image segmentation developed by researchers at Meta. SAM gained prominence for its ability to enable zero-shot transfer to various segmentation tasks, having been trained on over 1 billion masks from 11 million diverse images. Due to its extensive training dataset, SAM demonstrates the capability to generalize to new tasks beyond those encountered during training. SAM comprises of three main components: (i) An Image Encoder built from the Vision Transformer (ViT) [51], which preprocesses high-resolution inputs and runs once per image. (ii) A Prompt Encoder embedding dense prompts (i.e., masks) using convolutions, which are then summed element-wise with the image embedding. (iii) Mask Decoder that efficiently maps the image embedding, prompt embeddings, and an output token to a mask. Focal and dice loss are employed during training. SAMed, a variant of SAM adapted for medical segmentation, is introduced in [210]. SAMed is derived from SAM by freezing the image encoder and adopting a low-rank-based fine-tuning strategy (LoRA) [73]. This strategy approximates the low-rank update of the parameters in the image encoder and fine-tunes the lightweight prompt encoder and the mask decoder of SAM. SAMed was evaluated on the Synapse multi-organ segmentation dataset, achieving remarkable results. Building on the success of SAMed, SAMedOCT was adapted from SAMed to address the challenges posed by the RETOUCH grand challenge.

3.3.2 IAUNet_SPP_CL

IAUNet_SPP_CL, a combination of a graph-theoretic method, a fully convolutional neural network (FCN), and curvature regularization loss function is presented in [200]. The graph-theoretic method is employed in the preprocessing stage to delineate layers and regions of interest (ROI), while the FCN is utilized for fluid segmentation, employing the standard attention UNet as the backbone. The authors

enhanced the architecture by introducing spatial pyramid pooling (SPP) modules with four pooling maps at different scales in parallel, concatenating the original input after bilinear interpolation to enhance the network's capability to segment multi-scale objects. The curvature regularization loss function is applied to smooth boundaries and eliminate unnecessary holes within the predicted fluid lesions.

3.3.3 SFU

The SFU, a 3-part CNN-based and Random Forest (RF) framework is developed by [118]. The first part of the framework is used for pre-processing of the images, the second part consists of a 2D UNet architecture for the extraction of features and a RF classifier to classify the pixels at the third part. At the segmentation layer, axial motion between scans was corrected using cross-correlation by applying bounded variation 3D smoothing. This correction aimed to reduce the effect of speckle while preserving and enhancing the boundaries between retinal layers. To prevent overfitting during training, a dropout layer was introduced before the 1 to 1 convolutional layer. Additionally, to address data limitations, data augmentation techniques such as flipping, rotation, and zooming were applied during preprocessing.

3.3.4 UMN

The UMN, a combination of CNN and graph-shortest path (GSP) method is presented in [152]. The CNN is used for the segmentation of region of interest (ROI) and the GSP is further used for the segmentation of the layers and fluid from the ROI. B-scans were extracted from the 3D volumes for training. At the segmentation layer, the initial step involved segmenting the layers as ROI to efficiently detect the presence of fluids. Extracting the ROI helped reduce training time, as training the network on the entire image would be more time-consuming. The GSP was employed for pixel classification, mapping each pixel in the image to one node in the graph. Only local relationships between pixels were considered, and an 8-regular graph was constructed using the 8 neighbors of each pixel.

3.3.5 MABIC

The MABIC, a standard double-UNet architecture, is proposed in [93]. The method utilizes two UNet architectures connected in series, where the output of the first UNet serves as an input to the second UNet. The initial part takes raw images as input to extract the ROI. Additionally, in this initial part, dropout and maxout activation are applied at each layer to enhance accuracy and prevent overfitting. The subsequent part takes the extracted ROI and the segmentation mask as input. Importantly, there are no fully connected layers between encoding and decoding layers in the latter part.

3.3.6 RMIT

The RMIT, an approach using a combination of deep neural network and adversarial loss function is presented in [180]. The authors adapted the architecture from the standard UNet by incorporating a batch normalization layer in each block of convolutions. They introduced dropout at each skip connection to prevent overfitting and incorporated an adversarial loss function to estimate the loss during training.

3.3.7 RetinAI

The RetinAI, introduced in [10], is a standard 2D UNet with residual connections. The network was trained on B-scans. As part of the preprocessing, all the B-scans were normalized to the same resolutions, and horizontal flip, shear, rotation, shift, and Gaussian noise were applied for data augmentation. Categorical cross-entropy was used as the loss function during training.

3.3.8 SVDNA

A noise adaptation approach based on singular value decomposition (SVDNA) [99] is introduced as an unsupervised technique for noise transfer in the domain adaptation of retinal OCT images. The pipeline comprises of two phases. In the first phase, SVDNA is employed to generate masks, which are subsequently used to train a supervised segmentation network in the second phase. The model's performance was evaluated online, achieving a mean DS of 0.71 on the hidden test dataset. The authors didn't publish the AVD scores.

3.4 Experiments

3.4.1 Dataset

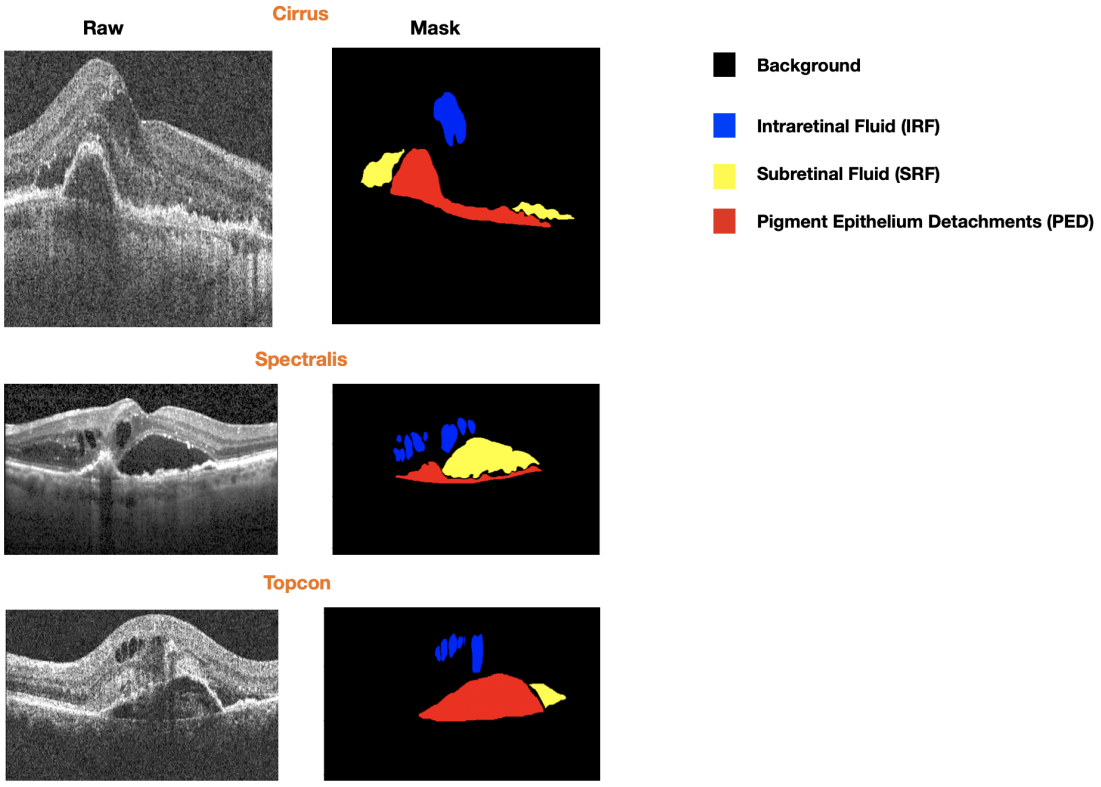

Figure 3.4: B-Scan examples of raw (column 1) and their corresponded annotated mask (column 2) of OCT volumes taken from the 3 device vendors (rows): Cirrus, Spectralis and Topcon. The classes are coloured as follows : Black for the background, blue for the Intraretinal Fluid (IRF), yelow for the Subretinal Fluid (SRF) and red for the Pigment Epithelium Detachments (PED).

The methods were validated on the MICCAI 2017 RETOUCH grande challenge dataset [21]. The dataset is publicly available and it consists of 112 OCT volumes of patients suffering with early AMD and DME collected from 3 device manufacturers: Cirrus, Spectralis and Topcon from 3 clinical centres : Medical University of Vienna (MUV) in Austria, Erasmus University Medical Centre (ERASMUS) and Radboud University Medical Centre (RUNMC) in the Netherlands. Examples of the dataset are shown in Figure 3.4.

The dimensions of the OCT volumes per vendor machine are as follows : Each volume of the Cirrus consists 128 B-Scans of 512×1024 pixels, Spectralis consists of 49 B-scans of 512×496 pixels and 128 B-Scans of 512×885 (T-2000) or 512×650 (T-1000) pixels for Topcon.

The training set consists of 70 volumes of 24, 24, and 22 acquired with Cirrus, Spectralis, and Topcon, respectively. Both the raw and annotated mask of the training set are made available to the public. The testing set consists of 42 OCT volumes of 14 volumes per device vendor. The raw or input of the testing set is available publicly but their corresponding annotated masks are held by the organiz-

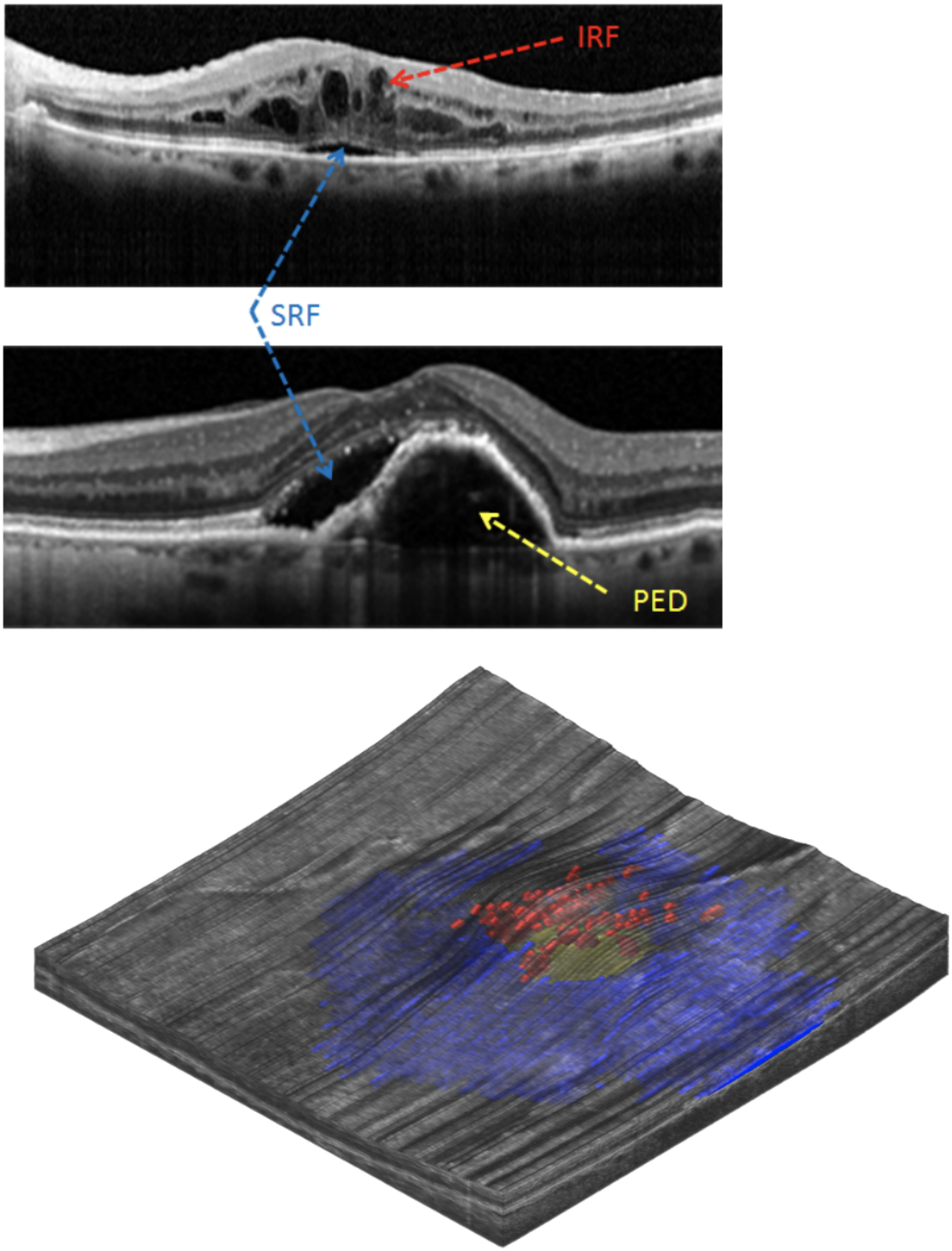

Figure 3.5: The three fluid types on an OCT slice (B-scan): Intraretinal Fluid (IRF) in red, Subretinal Fluid (SRF) in blue, and Pigment Epithelium Detachment (PED) in yellow. Volume rendering of different fluids inside the retina. Each subfigure represents a different patient. Image taken from [21]

ers of the challenge. Submission and evaluation of prediction on the testing dataset is arranged privately with the organizers and the results are sent to the participants.

Manual annotation was done by 6 grader experts from 2 medical centres : MUV (4 graders supervised by an ophthalmology resident), and RUNMC (2 graders supervised by a retinal specialist). The dataset is annotated for 4 classes of 1 background labelled as 0 and 3 fluids which are : Intraretinal Fluid (IRF) labelled as 1, Subretinal

Fluid (SRF) labeled as 2 and Pigment Epithelium Detachments (PED) labelled as 3.
Intraretinal fluid (IRF) consists of contiguous fluid-filled spaces containing columns of tissue. These spaces may appear as distinct hyporeflective cystoid pockets on OCT, and are sometimes referred to as intraretinal cystoid fluid.
Subretinal fluid (SRF) is the accumulation of clear or lipid-rich exudate in the subretinal space, located between the neurosensory retina and the underlying retinal pigment epithelium (RPE).
Pigment Epithelial Detachment (PED) is specific to AMD and involves the detachment of the retinal pigment epithelium (RPE), along with the overlying retina, from the Bruch's membrane due to fluid accumulation. PED can present as three subtypes: serous, fibrovascular, or drusenoid, all of which are considered and annotated as PED in the challenge.
A demonstration of these fluids are shown in Figure 3.5 and 3.4

The RETOUCH dataset is particularly interesting because of its high level of variability. It was collected using multiple device vendors, the sizes and number of B-Scans varies per device vendor, and it was collected and annotated in multiple clinical centres. Also, for fair comparison the annotated testing set is held by the organizers and submission is curbed to a maximum of 3 per participating team.

3.4.2 Training and Testing

Training was done on the 70 OCT volumes of the training set (both raw and mask volumes). The estimated probabilities and predicted segmentation of the testing set (42 raw volumes) were submitted to the challenge organizers for blinded evaluation on the ground truth or masks. The shapes of the input images were the same as the original image shape, as follows: $512 \times 1024 \times 128$ for Cirrus, $512 \times 496 \times 49$ for Spectralis and $512 \times 885 \times 128$ or $512 \times 650 \times 128$ for Topcon. nnUNet_RASPP leverages nnU-Net [84] self-parameterizing preprocessing techniques. Given the datasets, it extracts information such as modality, shape, and spacing (data fingerprint). Based on hardware constraints (GPU memory), hyperparameters such as network topology, image resampling methods, and input patch sizes are determined. During training, other hyperparameters were fixed as follows: the learning rate was set to 0.01, with a maximum of 1000 training epochs. The loss function was a combination of Cross Entropy and Dice loss, optimized using ADAM. Data augmentation was done on the fly, including random rotations, random scaling, random elastic deformations, gamma correction, and mirroring. nnUNet_RASPP was trained for 14 hours on an NVIDIA RTX A5000 GPU workstation. The code was written in Python using the PyTorch library.

Also, to further evaluate the robustness and generalisability of the methods, the predicted segmentation of the algorithm was evaluated on OCT volumes from two vendor devices and tested on the third. In this case OCT volumes from the third vendor device weren't seen during training. For this experiment, two sets of weights were generated which are: (1) Training on 46 OCT volumes from both Spectralis (24 OCT volumes) and Topcon (22 OCT volumes) and evaluated on 14 OCT volumes from the Cirrus testing set and (2) training on 48 OCT volumes from both Cirrus (24 OCT volumes) and Spectralis (24 OCT volumes) and evaluated on 14 OCT volumes from the Topcon testing set. Again the same environmental settings were used to conduct all the experiments.

In the detection task the estimated probabilities of presence of each fluid type is plotted using the receiver operating characteristics (ROC) curve. The area under the curve (AUC) which measures the ability of a binary classifier to distinguish between classes is used as the evaluation matrice. The AUC gives a score between 0 and 1 with 1 being the perfect score and 0 is the worst.

For the segmentation task, two evaluation matrices are used to measure the performance of the algorithms:

1. The Dice Score (DS) [26, 178, 134] which is twice the intersection, divided by the union. It measures the overlapping of the pixels in the range from 0 to 1 with 1 being the perfect score and 0 being the worst.

2. The Absolute Volume Difference (AVD) [178] which is the absolute difference between the predicted and the ground truth. The value ranges from 0 to 1 with 0 being the best result and 1 being the worst.

The equation to calculate the DS is shown on Eqn 3.1 and that for AVD in Eqn 3.2. Where X is the raw input or raw image, Y is the ground truth, or mask, $\cap$ is the intersection and $||$ is the absolute value.

$$DS = \frac{2|X \cap Y|}{|X| + |Y|} \tag{3.1}$$

$$AVD = |X| - |Y| \tag{3.2}$$

We used the Friedman test, a non-parametric statistical test, to detect differences in performance between the teams/algorithms evaluated on the segment classes. We computed the Friedman test statistic to check for significant differences and ranked the algorithms per segment class. The formula to calculate the Friedman test statistic is provided in Eqn 3.3, that to compute the degrees of freedom in Eqn 3.4 and that to compute the ranking in Eqn 3.5. A high-level interpretation of the hypothesis is provided in 3.4.2.

Friedman Test Statistic:

We computed the Friedman test statistic as:

$$\chi_F^2 = \frac{12N}{k(k+1)} \sum_{i=1}^{k} R_i^2 - 3N(k+1) \tag{3.3}$$

where: $k =$ Number of algorithms, $N =$ Number of segment classes, $R_i =$ Average rank of algorithm i.

Degrees of Freedom:

We computed the degrees of freedom as:

$$df = k - 1 \tag{3.4}$$

where: df is the degrees of freedom, and k is the number of algorithms.

Ranking Computation:

For each segment class j, we ranked the algorithms i based on performance. The average rank for each algorithm across all classes is given by:

$$R_i = \frac{1}{N} \sum_{j=1}^{N} r_{ij} \tag{3.5}$$

where: R_i = Average rank of algorithm i, N = Number of segment classes, and r_{ij} = Rank of algorithm i on segment class j.

Hypothesis Interpretation:

Null Hypothesis H_0: This indicates there is no significant difference in the rankings of the algorithms.
Alternative Hypothesis H_1: This indicates that at least one algorithm performs significantly differently.

If $p < \alpha$ ($\alpha = 0.05$), we reject H_0 and conclude that there is a significant difference between the algorithms.

3.4.3 Results

In this section we report the performance for the detection task measured by the Area Under the Curve (AUC), and the segmentation task measured by the Dice Score (DS) and Absolute Volume Difference (AVD) for the nnUNet_RASPP, and baseline nnU-Net. We also compare our results to the current state-of-the-arts (SOTA) architectures.

The segmentation performance grouped by segment classes per algorithm measured in DS is illustrated in Table 3.1 with the corresponding diagram in Figure 3.6, and that measured in AVD is illustrated in Table 3.2 with corresponding diagram in Figure 3.7.

We employed the Friedman test to assess the statistical significance of the algorithms' performance based on the combination of all three metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC). The results, including algorithm rankings and scores, are presented in Table 3.6.
According to the performance results, we noticed the following:

1. The nnUNet_RASPP and nnU-Net outperform the current SOTA architectures by a clear margin with a mean DS of 0.823 and 0.817 respectively. Also, obtaining a mean AVD of 0.036 for nnU-Net and 0.041 for nnUNet_RASPP.

2. Enhancing the nnU-Net improved the performance. The SRF class was the most difficult to segment with nnUNet_RASPP (the enhanced version of nnU-Net) obtaining the best DS of 0.80 which is 2% higher than the standard nnU-Net and and 5% higher than the best SOTA architecture. The nnUNet_RASPP also obtained the best SRF AVD of 0.016 compare to 0.017 of the baseline nnU-Net or 0.026 of the best SOTA models.

3. The best mean AVD score of 0.032 is achieved by SAMedOCT.

4. The nnU-Net and nnUNet_RASPP possess the second and third-best mean AVD scores, but they exhibit better IRF (0.019 and 0.021 compared to 0.042) and SRF (0.017 and 0.016 compared to 0.020) AVD scores than SAMedOCT.

5. Apart from the IRF class, the nnUNet_RASPP has the best DS in every single class when compare to the other models/teams.

6. IAUNet_SPP_CL and nnUNet_RASPP jointly achieve the second-best mean AVD score of 0.036.

7. We observed that, overall, the CNN/DNN models exhibit slightly better performance than the foundational model (SAMedOCT). We believe this is because SAMedOCT is constructed with ViT as a backbone, and ViTs are more data-hungry than CNNs due to their ability to model long-range dependencies, as explained in [51].

8. The Friedman test on the combination of all 3 metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC) revealed statistically significant differences between at least two algorithms, with a p-values of 0.0099 and and a Friedman test statistic of 31.1602.

9. Based on the Friedman test rankings, nnU-Net was ranked first with the highest score, followed by nnU-Net_RASPP.

A detail break down of the DS and AVD per vendor device trained on the entire 70 volumes and tested on the holding 42 cases of the testing set is shown in Table 3.4 with the corresponding diagrams in Figure 3.9. We noticed the following:

1. nnUNet_RASPP outperformed the baseline nnU-Net and the state-of-the-arts models in two (Cirrus and Spectralis) of the 3 devices in both DS and AVD. The nnUNet RASPP model came in second place on the third device (Topcon) with a marginal difference from the baseline model, nnU-Net.

2. The nnUNet_RASPP and nnU-Net were the only two algorithms to maintain constant high level performance and generalisability across all classes and data sources in both DS and AVD. Both models constantly occupied the top 2 spots in performance per segment classes and vendor devices.

The generalization performance, measured in DS and AVD, is presented in Table 3.5 with its corresponding diagrams in Figure 3.11. It shows the results when trained on 2 vendor devices from the training set and tested on the third device from the holding testing set measured in DS and AVD. In this case because of the constraint of the evaluation submission (curb to 3 maximum per team) of the predicted segmentation on the testing set, results for nnU-Net are unavailable. Here we noticed that

1. nnUNet_RASPP outperformed the current SOTA architecture scoring a mean DS of 0.86 (10% higher than the second best) on the Cirrus device and 0.81 (6% higher than the second best) on the Topcon device.

2. nnUNet_RASPP also obtained the best AVD scores, scoring a mean of 0.0114 and 0.0878 on the Cirrus and Topcon devices respectively.

3. nnU-Net_RASPP still maintained its high level of robustness and generalisability with a consistently high level of performance measure in DS and AVD.

The RETOUCH online competition is still ongoing. At the time of writing, our nnUNet_RASPP is currently ranked first among 216 participants from both online and offline submissions. Details of the competition, including the leaders table, number of participants and other statistics, are availabe at: [1]. Also, we have made the source code of the implementation publicly available with free distribution under the Apache-2.0 license at [2]

The detection performance grouped by segment classes per algorithm measured by the AUC is illustrated in Table 3.3 with the corresponding diagram in Figure 3.8. Here the nnU-Net obtained a perfect AUC score of 1 for all three fluid classes and nnUNet_RASPP obtained an AUC score of 0.93, 0.97, and 1.0 for the IRF, SRF, and PED respectively.
The visualizations using orange arrows to highlight the fine details capture by nnUNet_RASPP when trained on two vendor devices from the training set and tested on the third from the training set are illustrated in Figure 3.13, 3.14, and Figure 3.15.

Methods	IRF	SRF	PED	Mean
nnUNet_RASPP	0.84	**0.80**	**0.83**	**0.823**
nnU-Net	**0.85**	0.78	0.82	0.817
SFU	0.81	0.75	0.74	0.78
SAMedOCT [56]	0.77	0.76	0.82	0.78
IAUNet_SPP_CL [200]	0.79	0.74	0.77	0.77
UMN	0.69	0.70	0.77	0.72
MABIC	0.77	0.66	0.71	0.71
SVDNA [99]	0.80	0.61	0.72	0.71
RMIT	0.72	0.70	0.69	0.70
RetinAI	0.73	0.67	0.71	0.70
Helios	0.62	0.67	0.66	0.65
NJUST	0.56	0.53	0.64	0.58
UCF	0.49	0.54	0.63	0.55

Table 3.1: Segmentation table of the Dice Scores (DS) by segment classes (columns) and teams (rows) for training on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

[1] https://retouch.grand-challenge.org/Results/
[2] https://github.com/ndipenoch/nnUNetRASPP.git

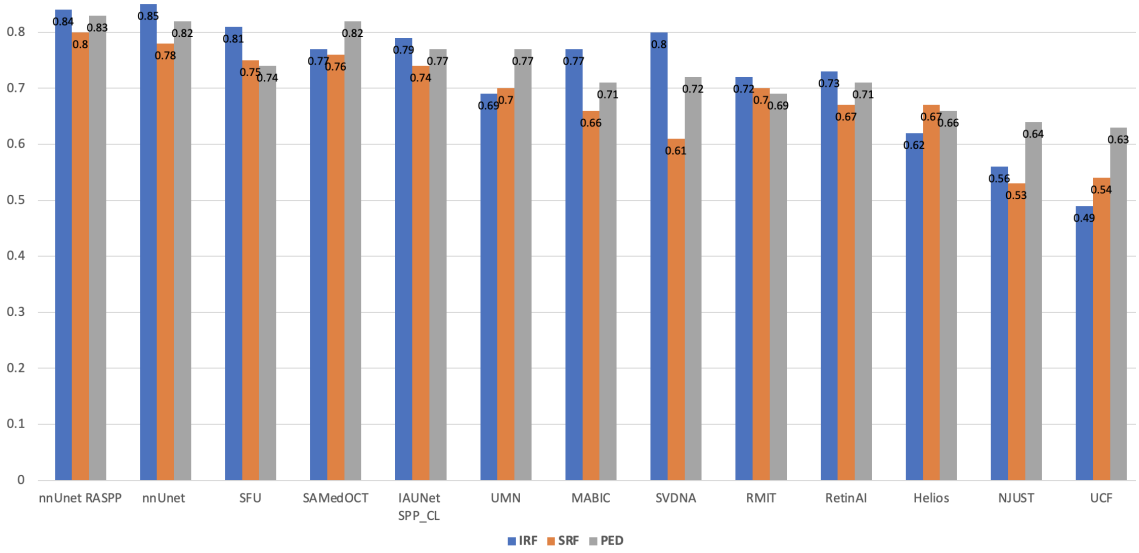

Figure 3.6: Performance comparison of segmentation measure in DS of the proposed methods: nnUnet_RASPP and nnU-Net, together with the current-state-of-the arts algorithms grouped by the segment classes when trained on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

Methods	IRF	SRF	PED	Mean
SAMedOCT [56]	0.042	0.020	**0.033**	**0.032**
nnU-Net	**0.019**	0.017	0.074	0.036
IAUNet_SPP_CL [200]	0.021	0.026	0.061	0.036
nnUNet_RASPP	0.023	**0.016**	0.083	0.041
SFU	0.030	0.038	0.139	0.069
UMN	0.091	0.029	0.114	0.078
MABIC	0.027	0.059	0.163	0.083
RMIT	0.040	0.072	0.182	0.098
RetinAI	0.077	0.041	0.237	0.118
Helios	0.051	0.055	0.288	0.132
NJUST	0.113	0.096	0.248	0.153
UCF	0.272	0.107	0.276	0.219

Table 3.2: Segmentation table of the Absolute Volume Difference (AVD) by segment classes (columns) and teams (rows) for training on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

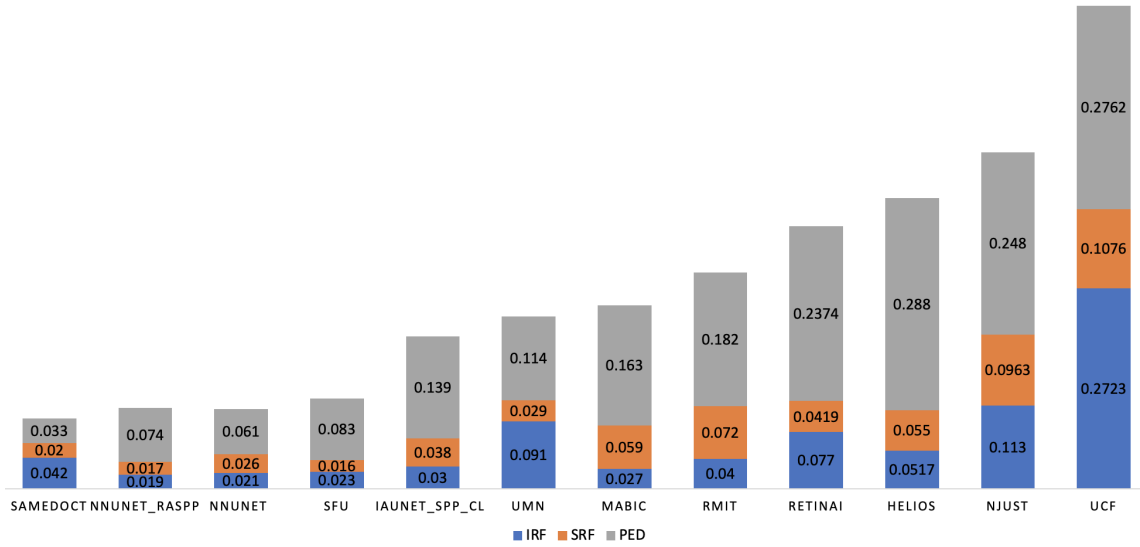

Figure 3.7: Performance comparison of segmentation measure in AVD of the proposed methods: nnUnet_RASPP and nnU-net, together with the current-state-of-the arts algorithms grouped by the segment classes when trained on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

Methods	IRF	SRF	PED	Mean
nnU-Net	**1.0**	**1.0**	**1.0**	**1.0**
SFU	**1.0**	**1.0**	**1.0**	**1.0**
nnUNet_RASPP	0.93	0.97	**1.0**	0.97
Helios	0.93	**1.0**	0.97	0.97
UCF	0.94	0.92	**1.0**	0.95
MABIC	0.86	**1.0**	0.97	0.94
UMN	0.91	0.92	0.95	0.93
RMIT	0.71	0.92	**1.0**	0.88
RetinAI	0.99	0.78	0.82	0.86
NJUST	0.70	0.83	0.98	0.84

Table 3.3: Detection table of the Area Under the Curve (AUC) by segment classes (columns) and teams (rows) for training on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

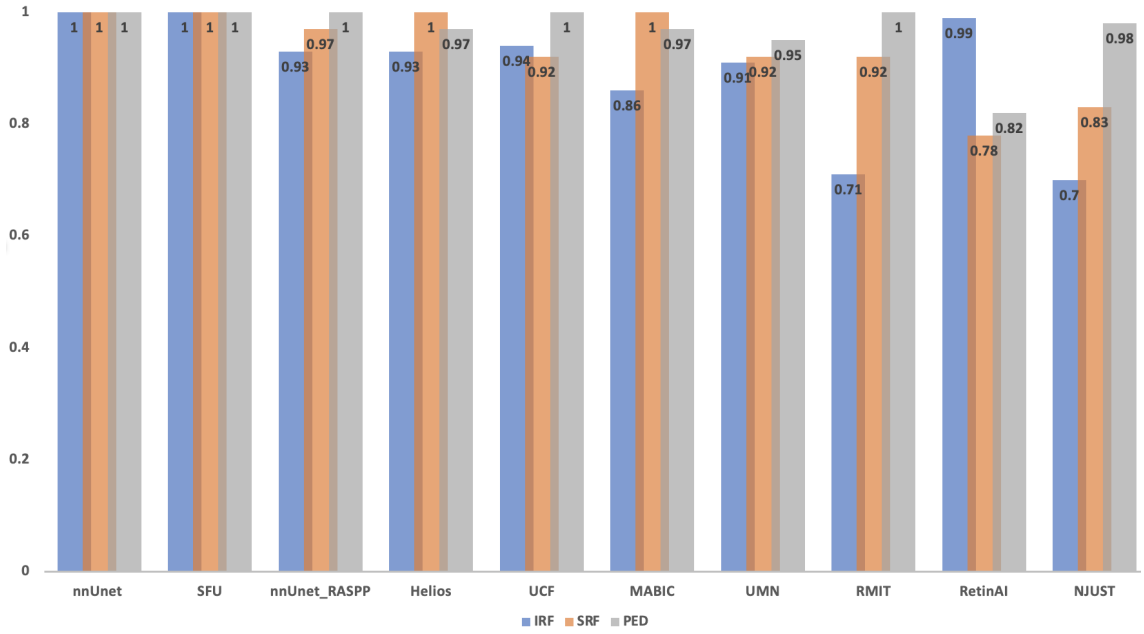

Figure 3.8: Detection performance comparison by DS of the nnU-Net_RASPP and baseline nnU-Net, together with the state-of-the-arts algorithms grouped by the segment classes when trained on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set.

Cirrus						
Methods	IRF		SRF		PED	
nnU-Net_RASPP	**0.91**	**0.00670**	**0.80**	**0.00190**	**0.89**	0.021700
nnU-Net	**0.91**	0.00850	**0.80**	**0.00190**	0.88	**0.02060**
SFU	0.83	0.020388	0.72	0.008069	0.73	0.116385
UMN	0.73	0.076024	0.62	0.007309	0.82	0.023110
MABIC	0.79	0.018695	0.67	0.008188	0.73	0.091524
RMIT	0.85	0.037172	0.64	0.005207	0.76	0.079259
RetinAI	0.77	0.046548	0.66	0.008857	0.82	0.040525
Helios	0.70	0.038073	0.66	0.008313	0.69	0.097135
NJUST	0.57	0.077267	0.55	0.024092	0.69	0.144518
UCF	0.57	0.174140	0.54	0.028924	0.66	0.215379
Spectralis						
Methods	IRF		SRF		PED	
nnUNet_RASPP	**0.89**	**0.030100**	0.68	**0.008400**	**0.81**	**0.068600**
nnUNet	**0.89**	0.031400	0.62	0.012600	0.80	0.073600
SFU	0.87	0.033594	**0.73**	0.020017	0.76	0.135562
UMN	0.76	0.072541	0.72	0.013499	0.74	0.121404
MABIC	0.83	0.036273	0.59	0.033384	0.75	0.181842
RMIT	0.69	0.121642	0.67	0.026377	0.70	0.228323
RetinAI	0.77	0.026921	0.65	0.036062	0.71	0.120528
Helios	0.61	0.030149	0.53	0.035625	0.63	0.330431
NJUST	0.60	0.080740	0.38	0.076071	0.52	0.412231
UCF	0.41	0.407741	0.31	0.155769	0.52	0.414739
Topcon						
Methods	IRF		SRF		PED	
nnU-Net_RASPP	0.72	0.032500	**0.93**	0.037800	**0.78**	0.157300
nnU-Net	**0.74**	**0.015900**	0.92	**0.036300**	**0.78**	**0.127700**
SFU	0.72	0.039515	0.80	0.085907	0.74	0.164926
UMN	0.59	0.125454	0.77	0.066680	0.76	0.197794
MABIC	0.68	0.025097	0.73	0.134050	0.65	0.215687
RMIT	0.63	0.072609	0.78	0.094004	0.60	0.404842
RetinAI	0.66	0.045674	0.70	0.171808	0.60	0.385178
Helios	0.56	0.086773	0.81	0.119888	0.65	0.435057
NJUST	0.52	0.181237	0.66	0.188827	0.70	0.187733
UCF	0.48	0.235298	0.76	0.134283	0.61	0.200602

Table 3.4: Segmentation table of the Dice Score (DS) and Absolute Volume Difference (AVD) by segment classes (columns) and teams (rows) for training on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set per device.

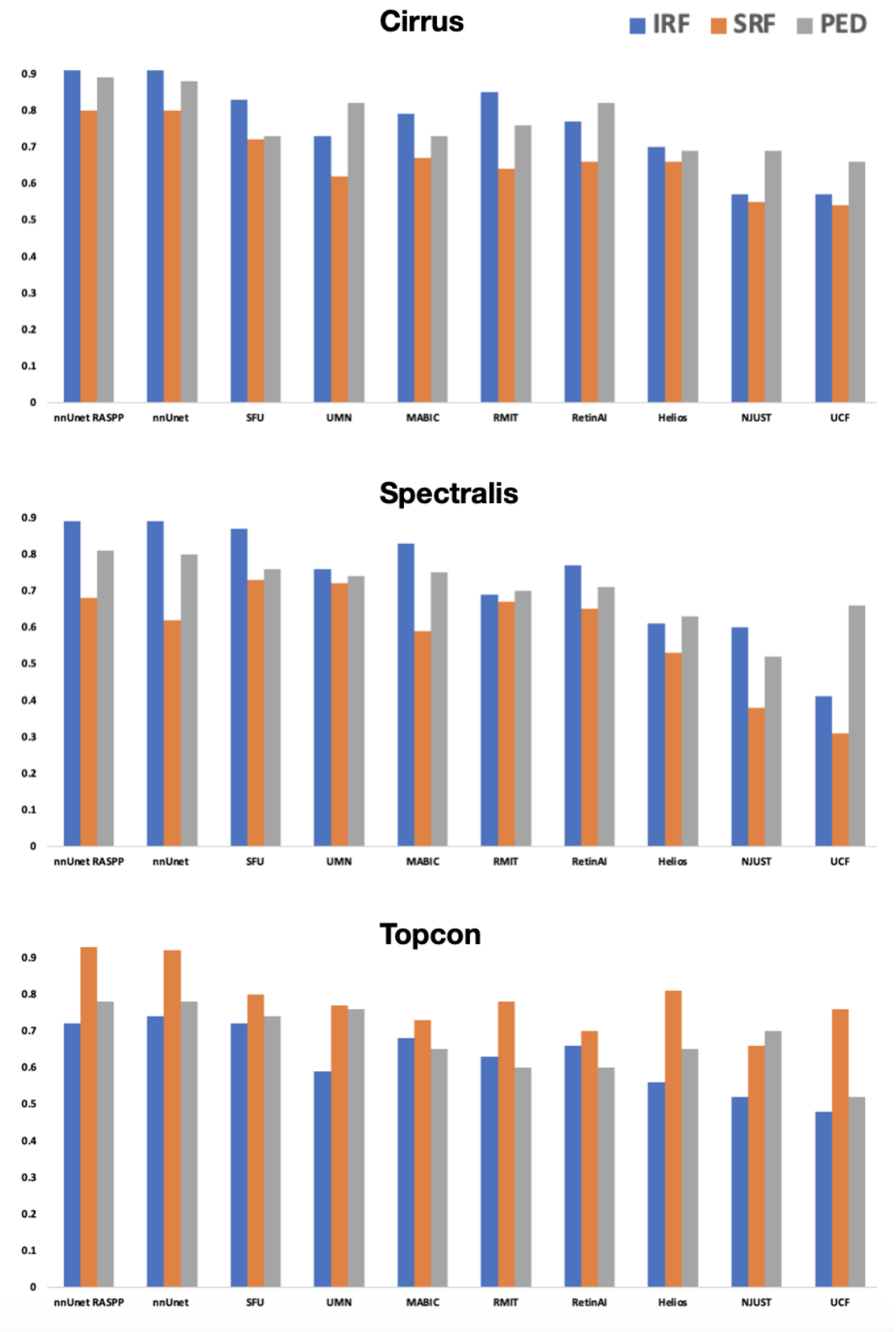

Figure 3.9: Performance comparison of segmentation measure in DS of the proposed methods: nnUnet_RASPP and nnU-net, together with the current-state-of-the arts algorithms grouped by the segment classes when trained on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set per device.

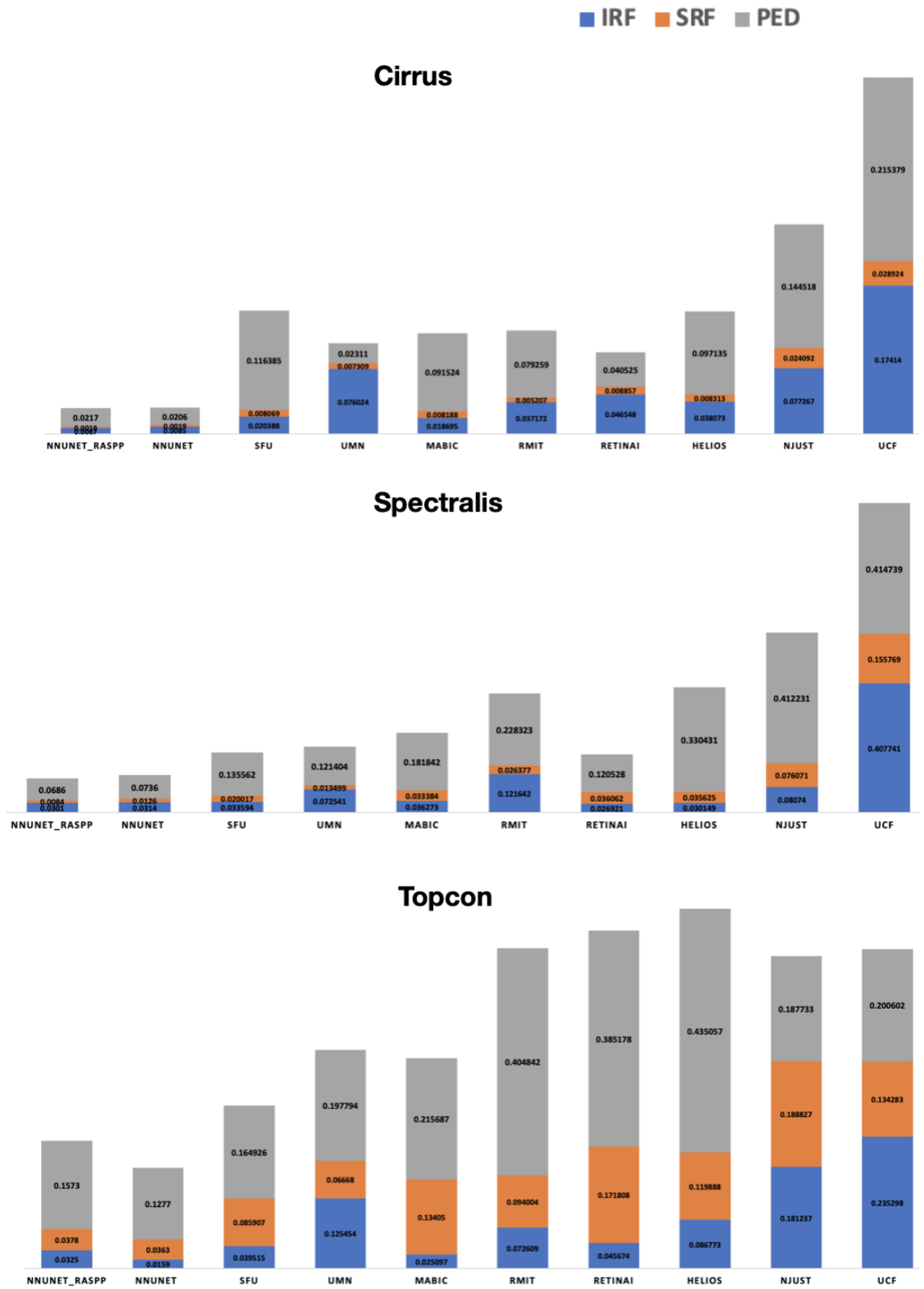

Figure 3.10: Performance comparison of segmentation measure in AVD of the proposed methods: nnUnet_RASPP and nnU-net, together with the current-state-of-the arts algorithms grouped by the segment classes when trained on the entire 70 OCT volumes of the training set and tested on the holding 42 OCT volumes from the testing set per device.

Cirrus								
Teams	IRF		SRF		PED		Mean	
nnUNet_RASPP	**0.90**	**0.0122**	**0.78**	**0.0031**	**0.89**	**0.019**	**0.86**	**0.0114**
SFU	0.83	0.0204	0.72	0.0081	0.73	0.1164	0.76	0.0483
UMN	0.73	0.0760	0.62	0.0073	0.82	0.0231	0.72	0.0355
MABIC	0.79	0.0187	0.67	0.0082	0.73	0.0915	0.73	0.0395
RMIT	0.85	0.0372	0.64	0.0052	0.76	0.0793	0.75	0.0406
RetinAI	0.77	0.0466	0.66	0.0089	0.82	0.0405	0.75	0.0320
Helios	0.70	0.0381	0.66	0.0083	0.69	0.0971	0.68	0.0478
SVDNA [99]	0.61	–	0.66	–	0.74	–	0.67	–
NJUST	0.57	0.0773	0.55	0.0241	0.69	0.1446	0.60	0.0820
UCF	0.57	0.1741	0.54	0.0289	0.66	0.2154	0.59	0.1395
Topcon								
Teams	IRF		SRF		PED		Mean	
nnUNet_RASPP	0.72	**0.0201**	**0.93**	**0.0298**	**0.78**	**0.2119**	**0.81**	**0.0873**
SFU	0.72	0.0395	0.80	0.0859	0.74	0.1649	0.75	0.0968
UMN	0.59	0.1255	0.77	0.0667	0.76	0.1978	0.71	0.1300
SVDNA [99]	0.61	–	0.80	–	0.72	–	0.71	–
MABIC	0.68	0.0251	0.73	0.1341	0.65	0.2157	0.69	0.1250
RMIT	0.63	0.0726	0.78	0.0940	0.60	0.4048	0.67	0.1905
RetinAI	0.66	0.0457	0.70	0.1718	0.60	0.3852	0.65	0.2009
Helios	0.56	0.0868	0.81	0.1199	0.65	0.4351	0.67	0.2139
NJUST	0.52	0.1812	0.66	0.1888	0.70	0.1877	0.63	0.1859
UCF	0.48	0.2353	0.76	0.1343	0.61	0.2006	0.62	0.1900

Table 3.5: Generalisation table of the DS and AVD by segment classes (columns) and teams (rows) trained on 48 OCT volumes from 2 device sources and evaluated on 14 OCT volumes from the testing set on the third device that wasn't seen at training.

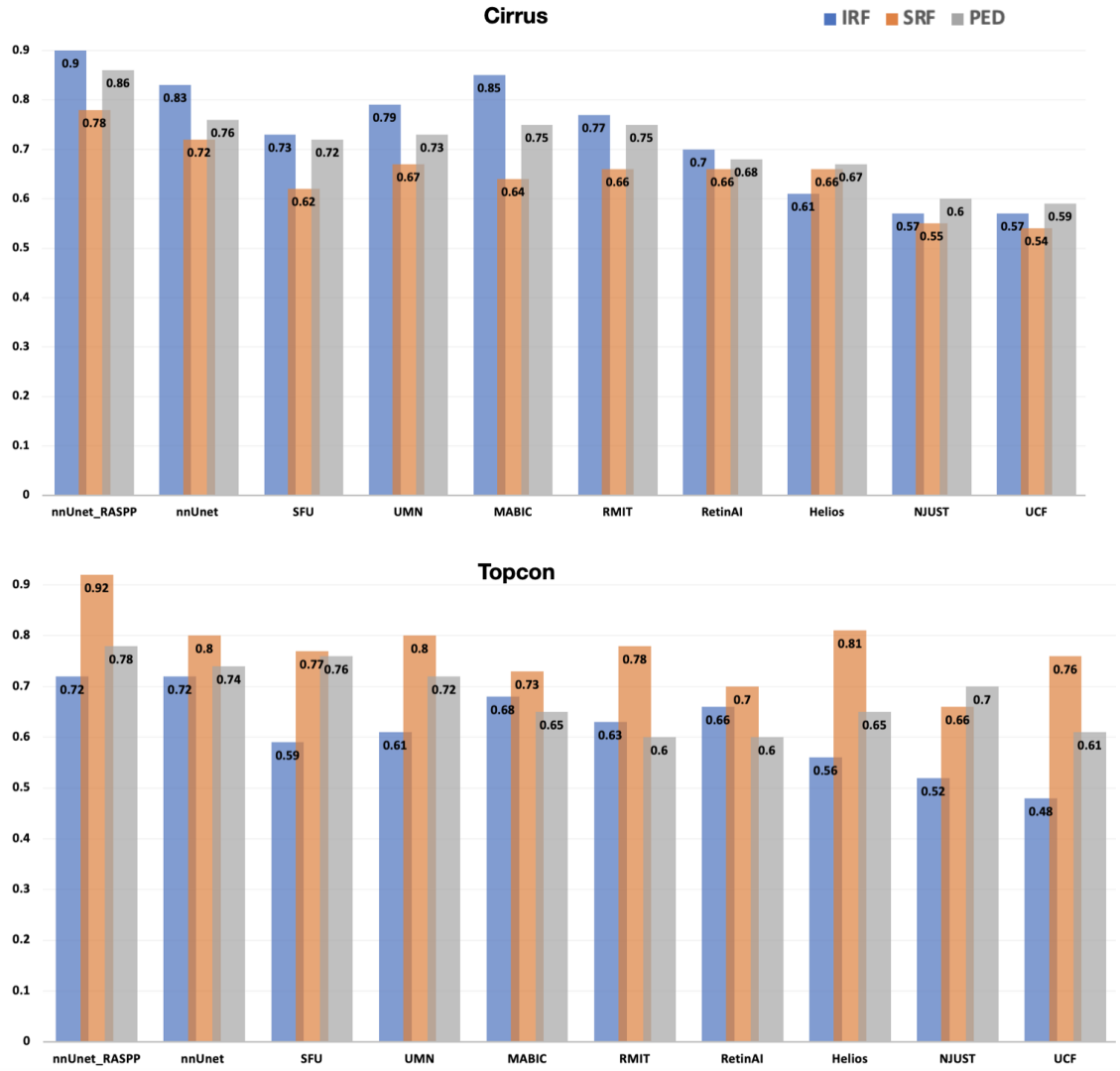

Figure 3.11: Generalisation performance comparison of segmentation measure in DS of the propose nnUnet_RASPP, together with the current-state-of-the arts algorithms group by the segment classes train on 46 OCT volumes from both Spectralis (24 OCT volumes) and Topcon (22 OCT volumes) and evaluated on the holding testing set (cirrus top and Topcon below).

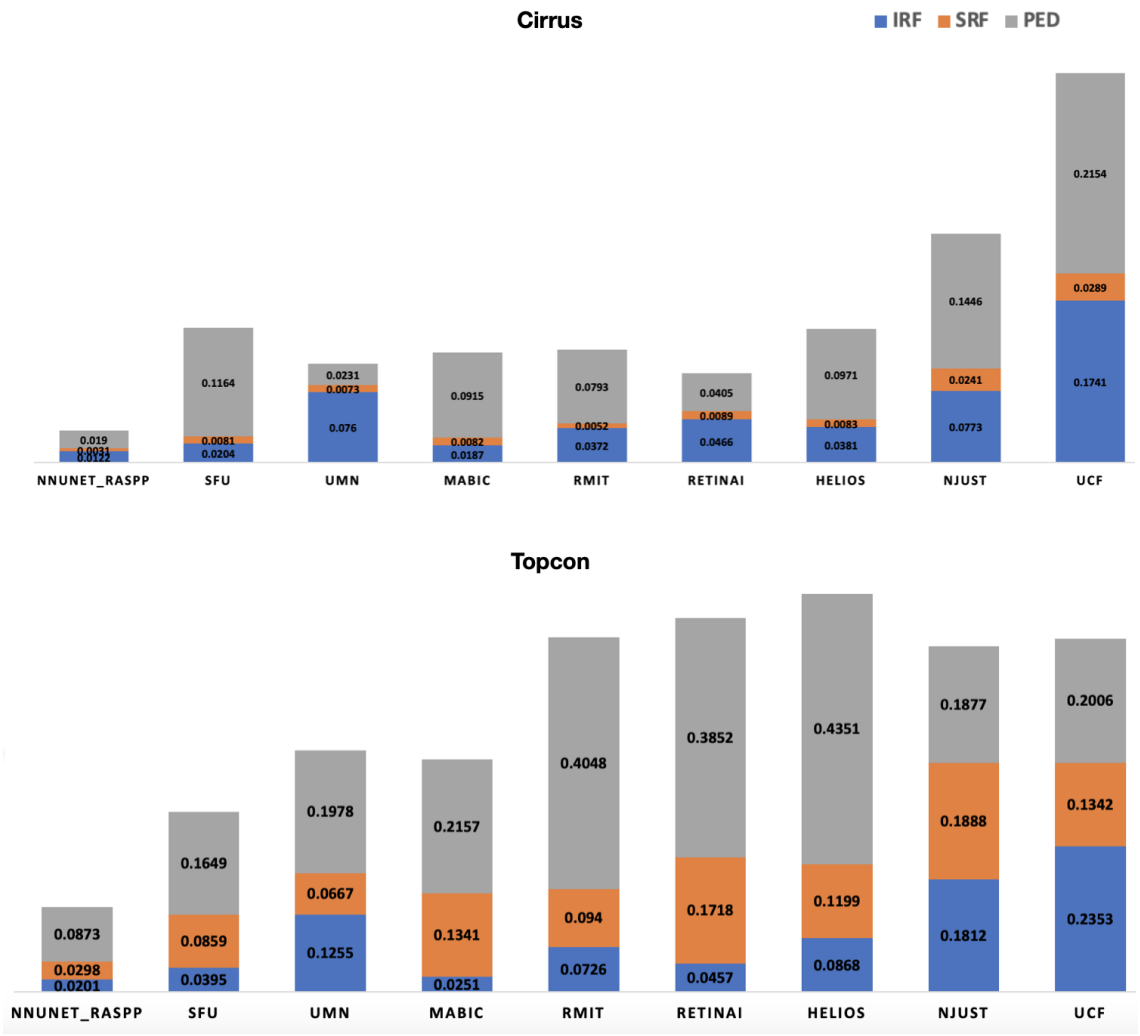

Figure 3.12: Generalisation performance comparison of segmentation measure in AVD of the propose nnUnet_RASPP, together with the current-state-of-the arts algorithms group by the segment classes train on 46 OCT volumes from both Spectralis (24 OCT volumes) and Topcon (22 OCT volumes) and evaluated on the holding testing set (cirrus top and Topcon bottom).

Number of segment classes: 3
Number of algorithms : 10
Degrees of freedom: (9, 2)
Significance level (alpha): 0.05
p-value: 0.0099
Friedman statistic: 21.6951
Hypothesis: Alternative Hypothesis
Significant: There is a significant difference between at least two algorithms (p-value < 0.05).

Rank	Algorithm	Ranking Score
1	nnU-Net	1.33
2	nnUNet_RASPP	1.67
3	SFU	2.33
4	UMN	4.33
5	MABIC	4.67
6	Helios	5.67
7	RMIT	6.00
8	RetinAI	6.67
9	UCF	7.33
10	NJUST	8.33

Table 3.6: The ranking (from best to worst) of the teams/algorithms based on the combination of all 3 metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC), using the Friedman test (a non-parametric test) indicates a significant difference between at least two of the algorithms, with a p-value of $0.0099 < 0.05$ and a Friedman test statistic of 31.1602.

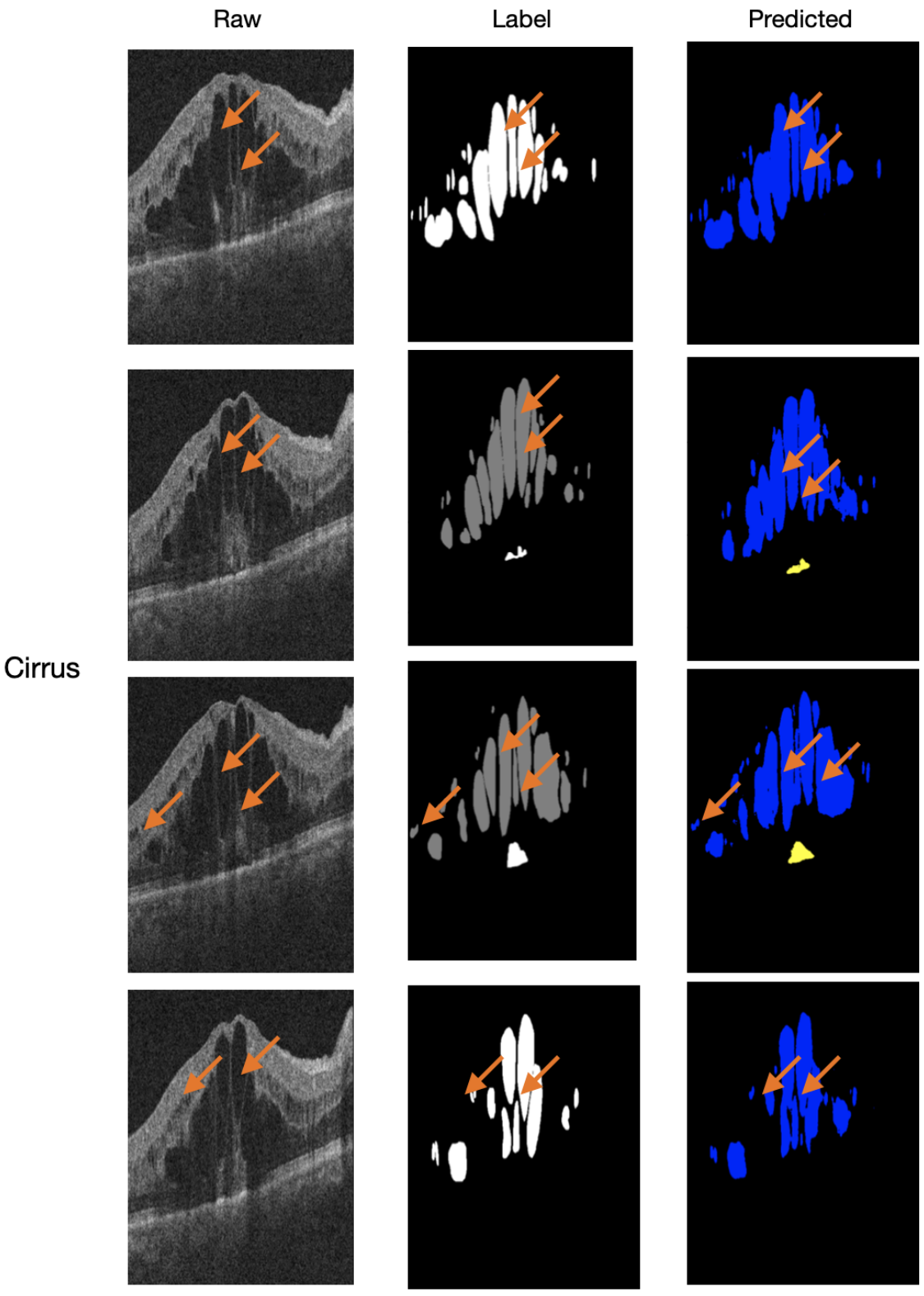

Figure 3.13: Examples of B-Scans to illustrate the visualization output/predicted of nnUnet_RASPP, in order of the raw/input, label/annotation and predicted/output in columns when trained on the training set of two vendor devices and tested on the training set of the third vendor device (Cirrus). Fine details capture by the model are indicated with orange arrows.

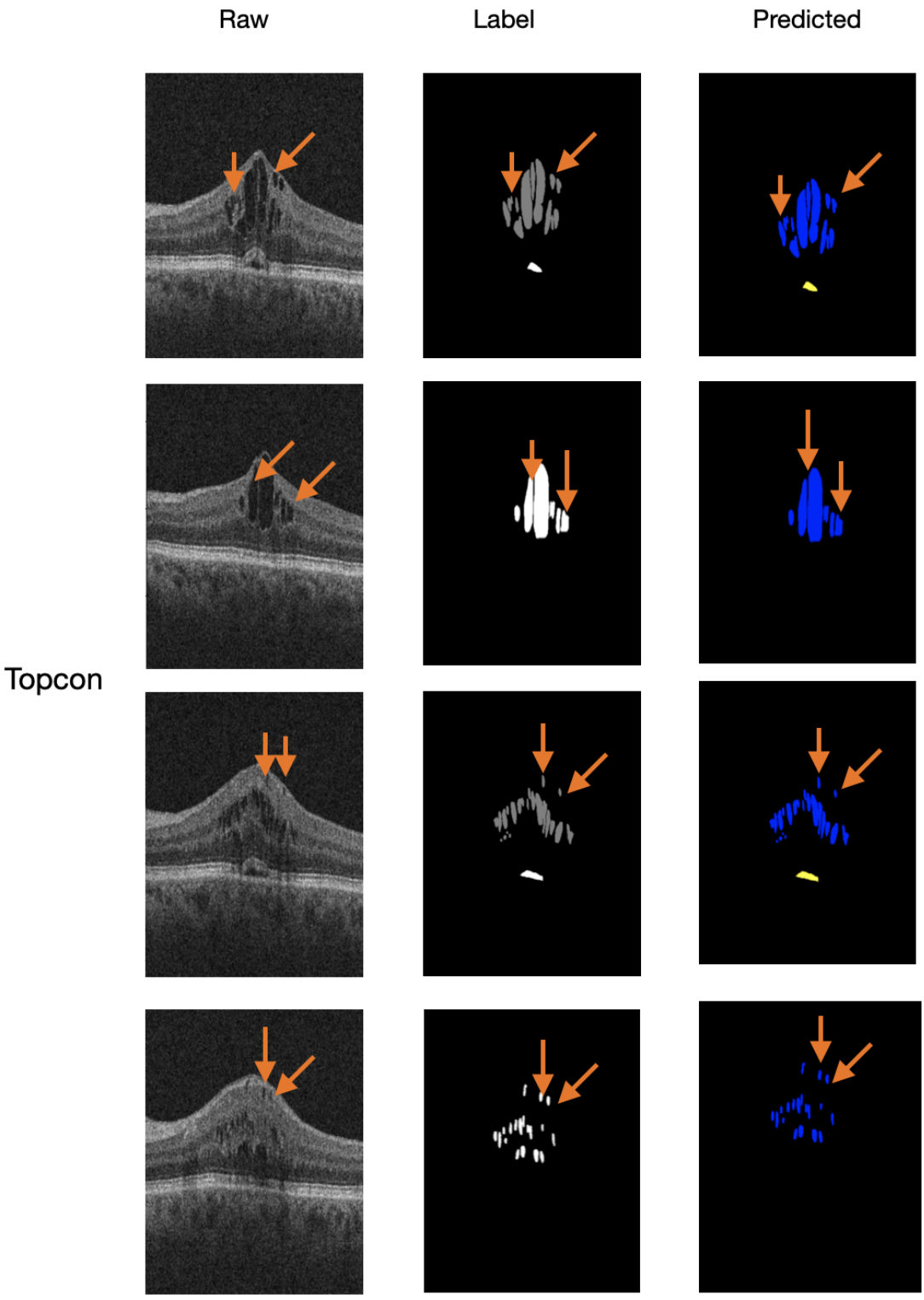

Figure 3.14: Examples of B-Scans to illustrate the visualization output/predicted of nnUnet_RASPP, in order of the raw/input, label/annotation and predicted/output in columns when trained on the training set of two vendor devices and tested on the training set of the third vendor device (Topcon). Fine details capture by the model are indicated with orange arrows.

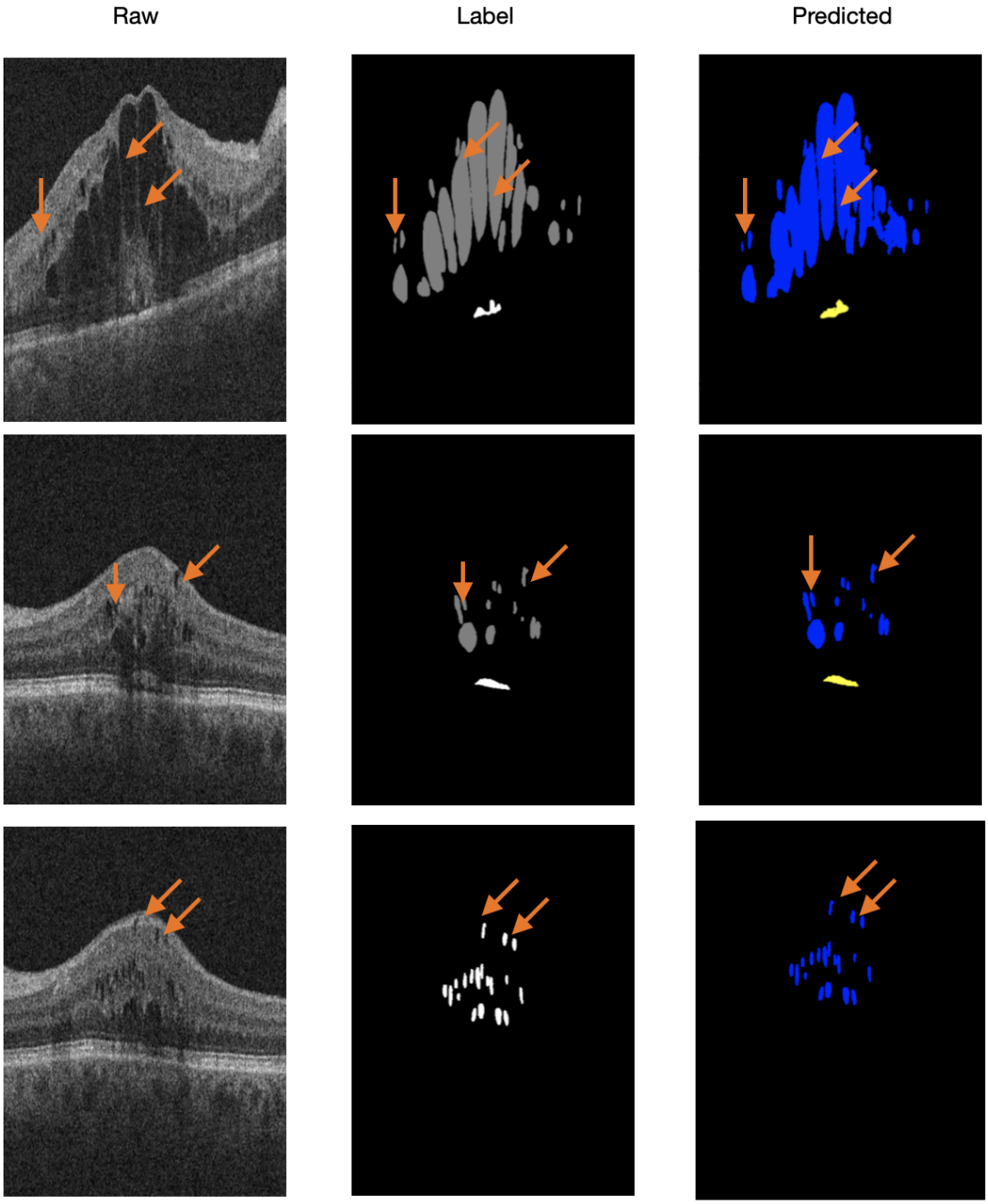

Figure 3.15: An example of a B-Scan to illustrate the visualization output/predicted of nnUnet_RASPP, in order of the raw/input, label/annotation and predicted/output when zoom out to highlights the fine details capture by the model using orange arrows. This is demonstrated when trained on the Spectralis training set and tested on Topcon, and vice versa.

3.5 Summary

In this chapter, we have investigated the problem of detection and segmentation of multiple fluids in retinal OCT volumes acquired from multiple device vendors. We improved the segmentation and generalization performance by enhancing the standard nnU-Net, to develop a novel algorithm called nnUnet_RASPP by incorporating an Atrous Spatial Pyramid Pooling (ASPP) block to : (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. Also, we added residual blocks to address overfitting, thereby improving the model's generalization performance. These enhancements address the challenges of high image variability associated with diverse, multi-source datasets, such as that of the Retouch challenge. As a result, nnUNet_RASPP demonstrates improved robustness and a greater ability to generalize across datasets with varying anatomical structures and image quality. Additionally, we conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed

Both nnU-Net and nnUnet_RASPP were evaluated on the MICCAI 2017 RETOUCH challenge dataset. We submitted predictions for both architectures and experimental results on the hidden test set show that the nnUnet_RASPP outperformed the current state-of-the-arts architectures and baseline nnU-Net by a clear margin as it occupy the first place of the challenge. Further more we are ranked first in the RETOUCH challenge with an overall best performance for both the online and offline results.

Our main contributions are as follows: (i) We enhanced the nnU-Net architecture [84] by incorporating an Atrous Spatial Pyramid Pooling (ASPP) block [35] at the input layer and residual blocks within the network's architecture. These modifications address the challenges of high variability in image quality, thereby improving robustness and generalization across diverse, multi-source datasets for this specific problem. (ii) We conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed.

The propose algorithms provide useful information for further diagnosis and monitoring the progress of retinal diseases such as AMD, DME and Glaucoma.

Chapter 4

Dynamic Network for Global Context-Aware Disease Segmentation in Retinal Images Using Multiple ASPP and SE Blocks

In Chapter 3, we explored the potential of using an Atrous Spatial Pyramid Pooling (ASPP) block [35] to capture global contextual information at the input layer, just before the down-sampling path. Given the high variability of diseases and image quality, in this chapter, we aim to improve segmentation in retinal images by using a dynamic network with ASPP blocks [35] at multiple locations: input, bridge, and output layers along with Squeeze-and-Excitation (SE) blocks [74] and a dense layer. Integrating multiple ASPP blocks alongside SE blocks enables the model to capture global contextual features at varying rates and across different locations within the network, spanning multiple resolutions and depths. This approach captures more detailed and comprehensive spatial dependencies in the input image, significantly enhancing the model's segmentation accuracy and generalization performance. We propose a novel algorithm termed Deep_ResUNet++, which enhances the ResUNet backbone by incorporating multiple ASPP blocks at the input, bridge, and output layers, along with Squeeze-and-Excitation (SE) blocks and a dense layer to capture global contextual information while dynamically adjusting the kernels size and depth of the network depending on the input image size. Deep_ResUNet++ provides an automated solution for the simultaneous segmentation of retinal layers and fluid regions in OCT scans. It integrates three key components into the network's backbone: (i) multiple ASPP and SE blocks to effectively capture global context, (ii) a dense layer for pixel segmentation and further capturing global information, and (iii) a mechanism to dynamically adjust the kernel size and depth of the network, enabling it to capture multi-scale information, fine details, and broader context. These innovations improve the network's overall performance and generalizability. Deep_ResUNet++ was evaluated on two benchmark datasets: the Annotated Retinal OCT Images (AROI) and the Duke DME datasets collected from patients suffering from two disease types. It demonstrated superior performance, surpassing the baseline ResUNet++ and state-of-the-art algorithms. On the AROI dataset,

it achieved a mean Dice score of 0.98, outperforming the second-best model by 0.01(1%), and consistently achieving over 0.90 across all classes. On the Duke DME dataset, it achieved a mean Dice score of 0.88, surpassing the second-best model by 0.02(2%). Deep_ResUNet++ demonstrates significant advancements in automated retinal OCT analysis, offering robust solutions for the diagnosis and monitoring of retinal diseases such as and Age-related Macular Degeneration (AMD) and Diabetic Macular Edema (DME).

The work presented in this chapter is published in [140], and [139]. These manuscripts are lead-authored by the author of this thesis, who made substantial contributions to the conception, data collection, processing, and writing, as well as sole contributions to the implementation and result analysis.

4.1 Introduction

In Chapter 3, we introduced Age-related Macular Degeneration (AMD) as the leading cause of severe vision impairment and blindness. Another eye diseases that do manifest in the retina is Diabetic retinopathy (DR). DR is a disease that damages the blood vessels in the retina, and it is the leading cause of blindness among working-aged adults in the United States [42]. Approximately 21 million people affected by DR also develop diabetic macular edema (DME) [22]. DME results from the accumulation of fluid in the macula, the central part of the retina where vision is sharpest, due to prolonged high blood sugar levels.

Currently, an effective treatment for these retina diseases is available in the form of anti-vascular endothelial growth factor (anti-VEGF) therapy [172], [23]. However, the effectiveness of this treatment depends on early diagnosis and frequent monitoring of the disease's progression, as this allows ophthalmologists to advise patients on behavioral changes such as diet change and doing regular exercise, which can help slow down the progression and, in some cases, prevent the disease from moving to later and more severe stage. Additionally, anti-VEGF drugs are expensive and require regular administration thus posing a soci-economic burden to both the patient and the healthcare system. Monitoring the progress of these diseases is crucial, however, the process is mostly manually done, which is time-consuming, labor-intensive, and prone to errors. Therefore, there is a need for an automated tool to diagnose and monitor retinal morphology and fluid accumulation accurately.

Optical Coherence Tomography (OCT) is a high-resolution, non-invasive imaging modality that provides qualitative information and visualizations of the retinal structure by acquiring a series of cross-sectional slices (B-scans). Developing an automated method to study the retina's anatomy from OCT B-scans and evaluate eye conditions like DME and AMD would be highly valuable.

To address these issue, we present a novel algorithm Deep_ResUNet++ developed in this chapter for simultaneously segmenting layers and fluids in retinal OCT B-scans. Unlike the common approach of treating retinal layers and fluid regions separately, this approach aim to provide an automatic solution for the simultaneous segmentation of both.

The rest of the chapter is organized as follows. A brief review of the previous studies is provided in Section 4.2. The proposed method is presented in 4.3. The experiments and result analysis are presented in 4.4. Finally, the summary is presented

in Section 4.5.

4.2 Background

The segmentation of retinal images has been a topic of great interest for several decades. Various methods have been explored to tackle this problem, ranging from traditional approaches such as graph-cut [164], [165], Markov Random Fields [163], [189] and level set methods [50], [49] to more recent deep learning techniques
Optical Coherence Tomography (OCT) is the current image-guided standard to analysis, diagnose and monitor the pathological changes in the retina. OCT was developed in the 1990s [77], but it only became commercially available in 2006. It allows for fast image acquisition and successful quantitative analysis due to its high quality and resolution. Some of the earliest segmentation approaches for retinal images include:Segmentation of retinal layers in OCT images using the graph method [62], segmentation of fluid in the retina in patients suffering from Macular Edema (ME)[1], and segmentation of fluid using the active contours approach [58]. Some of the recent approaches to segment retinal diseases in OCT images includes [21] for the segmentation of 3 fluids in retinal OCT images, [160] for the segmentation of retinal layers and fluids, [56] a large foundation model for the segmentation of three retinal diseases and many more.

4.3 Method

In this section, we introduce our proposed novel architecture, Deep_ResUNet++, an enhanced version of ResUNet++ [87]. We detail the modifications we made, their functionalities, how Deep_ResUNet++ differs from the original ResUNet++, and how the modifications would improve the model's generalization ability. We integrated three key innovations into the network's backbone: (1) An Atrous Spatial Pyramid Pooling (ASPP) blocks at multiple locations (input, bridge, and output layers) to: (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. (2) A dense layer at the classification layer to further improve the network's ability to capture global contextual features and for pixels classification. (3) A mechanism to dynamically adjust the kernel size and depth of the network based on the input image size, allowing the network to adapt to different feature scales, capturing both fine details and broader context. In the following subsections, we will provide a detailed outline of the components of Deep_ResUNet++. A high-level diagram demonstrating the architecture is shown in Figure 4.1.

4.3.1 Deep_ResUNet++

Encoding and Decoding Paths

The Deep_ResUNet++ model is fundamentally similar to the 2D U-Net architecture [157], featuring both an encoding and decoding path. The encoder phase captures

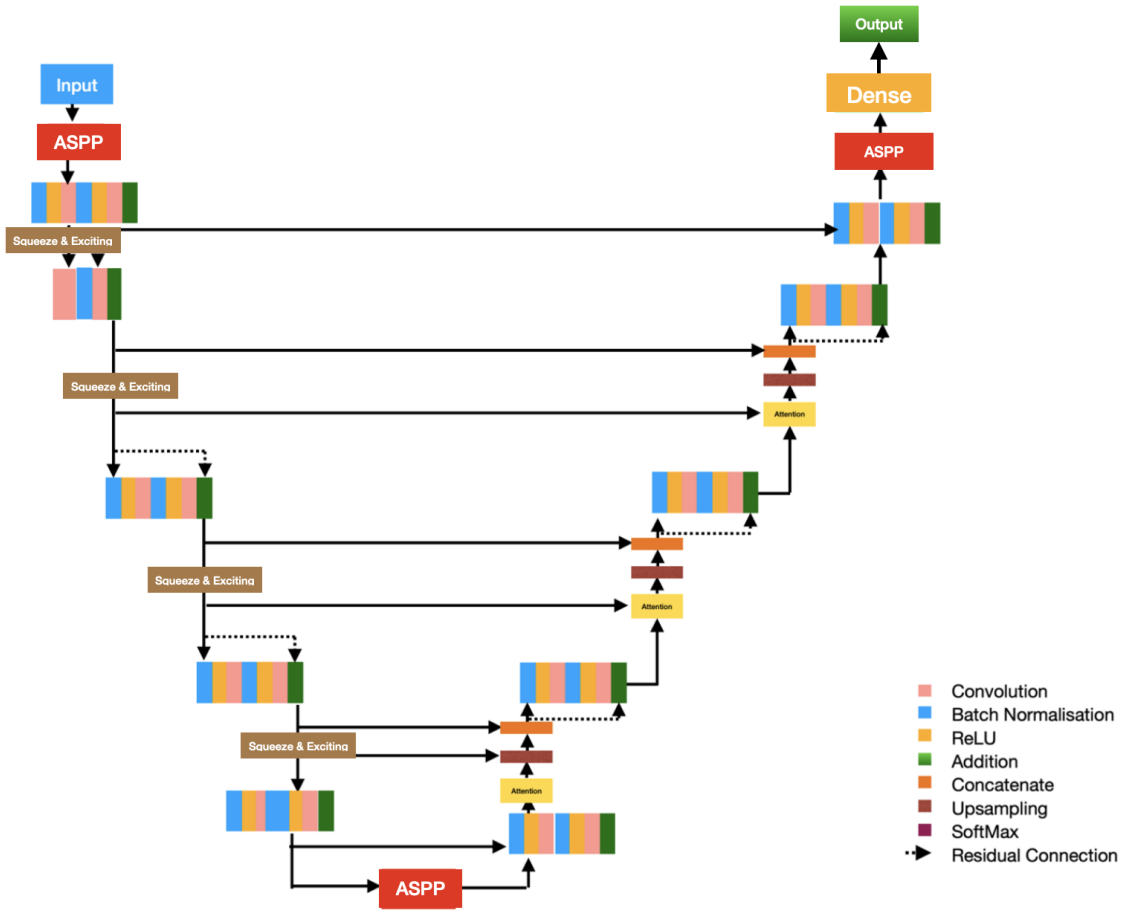

Figure 4.1: Structure of Deep_ResUNet++ demonstrating the Atrous Spatial Pyramid Pooling (ASPP) blocks along with Squeeze-and-Excitation (SE) to capture global features and the dense layer for pixel classification at the classification layer.

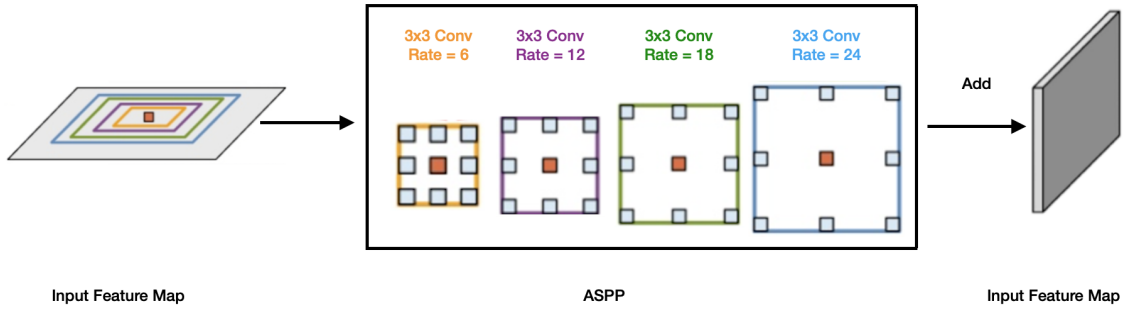

Figure 4.2: The ASPP captures global information by using multiple parallel filters with varying dilation rates.

local contextual information, while the decoder phase enables precise pixel localization, with a bridge layer connecting the two phases.

The encoder consists of convolutional blocks, each containing three sequential layers: batch normalization, ReLU activation, and convolution layer. Batch normalization helps prevent overfitting during training. Instead of the fixed 3×3 square kernels used in ResUNet++, we employ a strategy that dynamically adjusts the kernels size to match the image dimensions enabling the network to adapt to different feature scales, capturing both fine details and broader context. Zero padding

is applied to ensure that the feature map dimensions remain consistent before and after convolution, and a stride of one is used to avoid overlapping in the feature map construction.

The decoder also comprises of convolutional blocks consisting of, a batch normalization, ReLU activation, convolution layer, upsampling, and concatenation. The upsampling layer captures spatial information from the feature map, while the concatenation layer combines images from the encoder phase with their corresponding decoder phase, ensuring that the input and output image sizes match. The first three layers in the decoder follow the same setup as in the encoder.

The Bridge Layer

An Atrous Spatial Pyramid Pooling (ASPP) block serves as a bridge between the encoding and decoding phases. ASPP is an upsampled filtering technique used to capture global information within a feature map. It consists of multiple parallel atrous convolutional layers with different dilation rates. These blocks are designed to perform convolution with upsampled filters, allowing them to capture global contextual features efficiently while maintaining computational efficiency.

In contrast to ResUNet++, which employs three parallel filters with dilation rates of 6, 12, and 18, the Deep_ResUNet++ utilizes four parallel filters with dilation rates of 6, 12, 18, and 24. This adjustment accommodates the increased image capacity of the model, enhancing its ability to capture global information. The inclusion of the ASPP block is crucial for this task due to the variability in fluid types present in B-Scans, with at least one fluid type often absent in some scans. The ASPP block used in the Deep_ResUNet++ architecture is illustrated in Figure 4.2.

Deep Residual Learning

In a neural network, adding more layers and using activation functions like Sigmoid which compresses large input values into smaller values between 0 and 1 can cause the gradients of the loss function to approach zero. This phenomenon makes the network difficult to train and is known as the vanishing gradient problem in deep learning. One effective way to address this issue is by using skip or residual connections. These connections allow some layers to bypass the activation functions during training, thereby reducing the extent to which derivatives are diminished. This approach is crucial in Deep_ResUNet++ because increasing its depth heightens the likelihood of encountering the gradient problem.

Squeeze and Exciting Block

The Squeeze-and-Excitation (SE) block [74] leverages Global Average Pooling (GAP) to capture global context by averaging spatial information across feature maps. In the encoding path, GAP blocks are placed between convolutional layers to extract global contextual information. SE blocks dynamically recalibrate channel-wise feature responses, amplifying the most relevant features while suppressing less significant ones. When incorporated at multiple locations within the network, SE blocks enhance feature representations at various stages, improving feature discrimination and overall segmentation performance.

Dense Layer

Unlike ResUNet++, which uses 2D convolutions in the final layer of the decoding phase just before the output layer, the Deep_ResUNet++ incorporates a Dense layer in this position. This modification allows the network to learn and integrate information from all preceding features, enhancing its ability to capture global context. This is particularly advantageous for our problem, as fluid regions exhibit significant variability and lack the consistency seen in retinal layers.

Classification Layer

The classification layer is tasked with determining the class for each voxel or pixel in the final feature map. The Deep_ResUNet++ uses the SoftMax activation function to classify each pixel or voxel in the input feature map, assigning it to one of the classes or labels.

Dynamic Network

Unlike ResUNet++, which uses fixed kernel sizes and network depth, Deep_ResUNet++ dynamically adjusts the kernel size and network depth, allowing the architecture to adapt and capture multi-scale features, fine details, and broader context based on the shape and size of the input image. For square images, a 3×3 kernel is used to match the image's symmetrical structure, while for rectangular images, a 7×3 kernel is applied to better capture elongated spatial features. The network depth is also adjusted based on dataset size. For datasets with fewer than 1,000 samples, the depth is set to 4, whereas for datasets exceeding 1,000 samples, the depth is set to 5 to enhance feature representation and learning capacity. Given the high variability of retinal diseases and image quality, this flexibility is particularly useful for detecting small abnormalities (diseases/fluids) as well as larger structures (layers) in retinal images, thereby improving the model's overall performance, generalizability, and robustness.

Hyperparameter Settings

The encoder path consists of convolutional blocks, each containing a 7×3 (can change depending of the shape of the input image) convolutional layer, Batch Normalization (BN), ReLU activation, max pooling, and padding is $(\text{kernel\_size} - 1)/2$ with a stride of 1. The decoder path mirrors the encoder path's structure but replaces max pooling with upsampling and includes concatenation with corresponding encoder features for precise reconstruction. The network's depth was set to a maximum of 5 (can vary depending on the size of the dataset), corresponding to an input size of 1024×512 pixels. The ASPP blocks consist of 4 parallel filters with dilation rates of 6, 12, 18, and 24. K-fold cross-validation was employed with the value of K set to 6, Categorical Cross-Entropy was the loss function used, with AdaBound as the optimizer, the initial learning rate was set to 0.001, and the batch size was set to 4. The maximum training epoch was set to 200 with early stopping to prevent overfitting.

In Chapter 3, a single ASPP block was employed, whereas in this chapter, multiple ASPP and SE blocks are integrated at various locations within the network. This configuration enables the model to capture global contextual features at varying

rates and across different locations, encompassing multiple resolutions and depths. Such an approach allows the network to improve the model's generalizability by: (i) effectively capturing structures of varying sizes within the images, (ii) adapting more effectively to diverse dataset characteristics, such as variations in resolution and noise, (iii) capturing both local and global context for a more comprehensive understanding of the data, and (iv) reducing the model's dependence on features from any single scale, thereby significantly improving segmentation accuracy and generalization performance.

4.4 Experiments

Deep_ResUNet++ was evaluated on two benchmark public datasets: the Annotated Retinal OCT Images (AROI) [131] and the Duke DME [40] datasets. The following subsections provides detailed information about these datasets.

4.4.1 Annotated Retinal OCT Images (AROI) Dataset

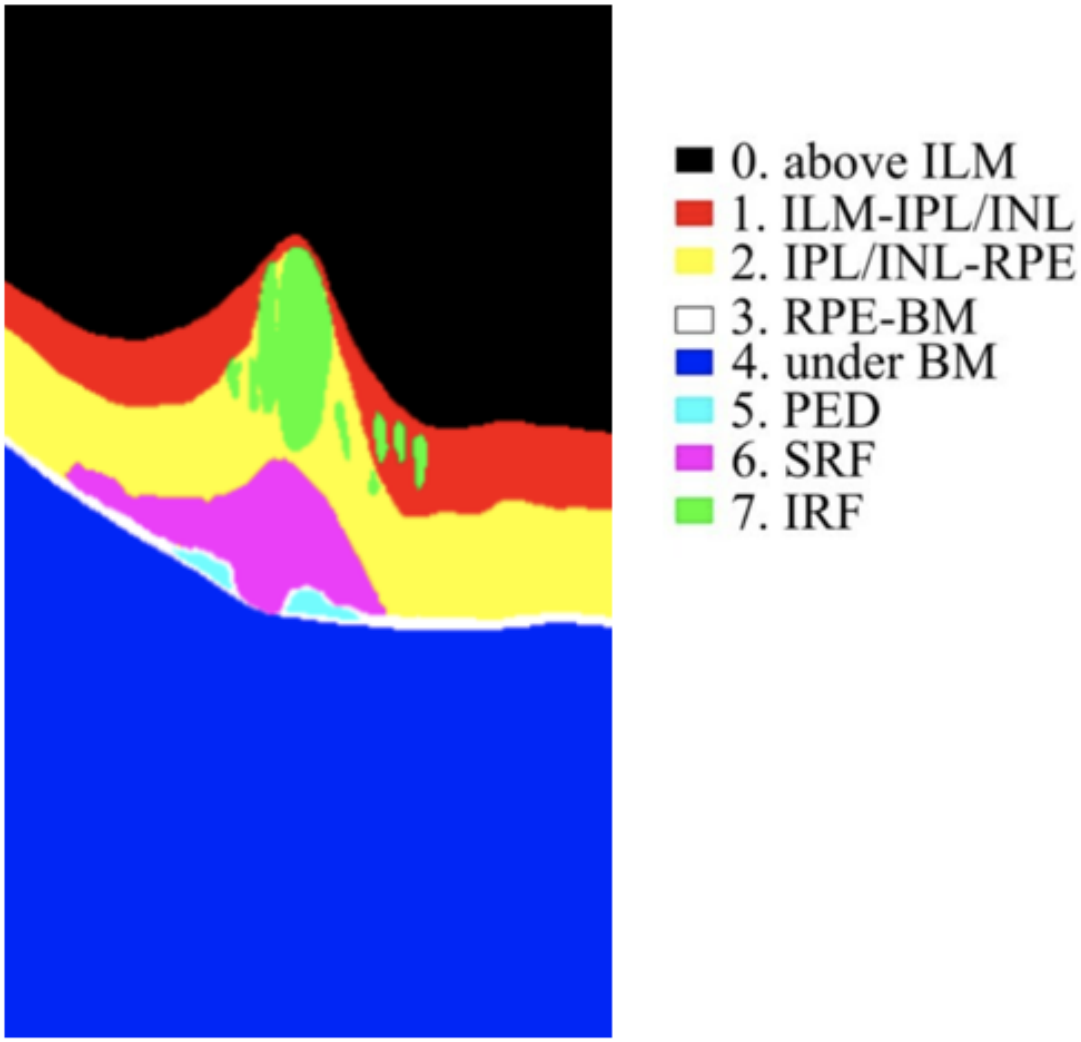

Figure 4.3: An example of annotation of the layers and fluids in the AROI dataset.

The Annotated Retinal OCT Images (AROI) database was collected using the Zeiss Cirrus HD OCT 4000. It consists of 128 B-scans per OCT image for each of the 25 patients with wet AMD, totaling 3,200 B-scans. Among these, 1,136 B-scans from 24 patients are annotated, and this subset was used for the experiments. The resolution of the B-scans is 1024×512 pixels, with a pixel size of 1.96×11.74 μm. In total, eight labels or classes were identified, and the number of labels per B-scan depends on the presence or absence of fluids. Subretinal fluid or subretinal hyperreflective material (SRF/SRHM) and Intraretinal fluid (IRF) are not present in all OCT volumes. SRF is absent in patients 13, 17, and 19, while SRF/SRHM are absent in patients 3, 4, 6-9, 17, 20-22, and 24.

The B-scans were labeled based on three categories: layer, fluids, and background. Historically, the retinal is consists of 10 layers, but for simplicity, these layers are grouped into three distinct classes: 1) Internal Limiting Membrane (ILM): Which is the area between the ILM and the Inner Plexiform Layer (IPL)/Inner Nuclear Layer (INL) boundaries, 2) Inner Plexiform Layer and Inner Nuclear Layer (IPL/INL): which is the area between the IPL/INL and the Retinal Pigment Epithelium (RPE) boundaries, and 3) Retinal Pigment Epithelium/Bruch's Membrane Complex (RPE/BM) which is the area between the RPE and BM boundaries.

Four main retinal fluids were identified and categorized into three classes: 1) Intraretinal Fluid (IRF), 2) Subretinal Fluid (SRF) and Subretinal Hyperreflective Material (SRHM), which are grouped together as SRF since they are located in the same area, and 3) Retinal Pigment Epithelial Detachment (PED).

Two background categories were identified: The area above the Internal Limiting Membrane (ILM), and The area below the Bruch's Membrane (BM). The classes are color-coded as follows: Black represents the area above the ILM, Red denotes the ILM layer, Yellow indicates the area between the IPL and INL layers, White is used for the RPE and BM layers, Blue denotes the area under the BM, Light Blue represents the PED fluid, Pink represents the SRF/SRHM fluids, and Green signifies the IRF. An example of the labeling and annotation of retinal layers and fluids is shown in Fig. 4.3.

4.4.2 Duke DME Dataset

The Duke DME dataset [40], consists of 110 B-scans from 10 patients with severe DME pathology and was collected using the standard Spectralis (Heidelberg Engineering, Heidelberg, Germany). The volumetric scans have a configuration of 61 B-scans and 768 A-scans, with an axial resolution of 3.87 µm/pixel, a lateral resolution ranging from 11.07 to 11.59 µm/pixel, and an azimuthal resolution ranging from 118 to 128 µm/pixel.

The images were annotated by two human experts across three categories (layer, fluid, and background), resulting in 10 classes: 1 fluid, 2 backgrounds, and 7 layers. Traditionally, retinal OCT includes 10 layers; however, for clarity, these layers are grouped into 7 distinct classes in this dataset: Inner Limiting Membrane (ILM), Nerve Fiber Layer to Inner Plexiform Layer (NFL-IPL), Inner Nuclear Layer (INL), and Outer Plexiform Layer (OPL).

The fluid class was identified, and the two background classes are the area above and below the retina. In this work, the classes are annotated with the following colors: Black for the area above and below the retina, Light Green for the ILM

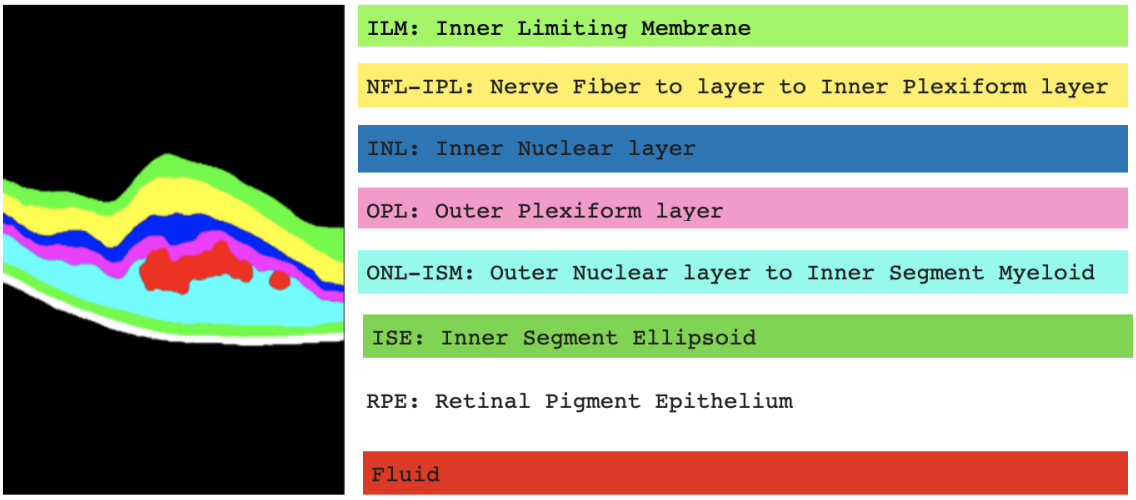

Figure 4.4: Annotation and labeling of the 10 segments (7 retinal layers, 2 backgrounds, and 1 fluid) in the Duke DME dataset.

layer, Yellow for the area between the NFL and IPL layers, Blue for the INL, Pink for the OPL layer, Light Blue for the area between the ONL and ISM layers, Green for the ISE layer, White for the RPE, and Red for the fluid.

An example of the annotation and labeling of classes is shown in Figure 4.4. It is important to note that the Duke DME dataset was collected for two specific problems: layer and fluid segmentation. Additionally, the fluid class exhibits high variability and is not present in some B-scans for certain patients, adding to the dataset's complexity.

4.4.3 Training and Testing

It is common practice to separate the segmentation of regular retinal layers from the detection of fluids, but in this work, we aim to perform both tasks simultaneously. K-fold cross-validation was used for training, validation, and testing for each dataset. The Dice score was the evaluation metric used to measure the performance of the algorithm. It is a similarity measure often used in the segmentation of medical images. The Dice score is the percentage of pixels or voxels in an image that are classified correctly per class or segment. It is calculated by taking twice the intersection and dividing it by the union for each class or segment as demonstrated in in Eqn (3.1). In this section, we will also refer to the proposed Deep_ResUNet++ as **Proposed**.

1) **Annotated Retinal OCT Images (AROI)**:
For the AROI dataset, to ensure a fair comparison, we used the same data splits as in the baseline model [130]. Each fold consists of B-scans from 4 patients. For example, the first fold includes patients 1, 2, 3, and 4, the second fold includes patients 5, 6, 7, and 8, and so on. Splitting B-scans from the same patient across training, validation, and test sets is not recommended, as adjacent B-scans are similar and could introduce bias. The test set comprises approximately 15% of the dataset. For all experiments, the parameters were set as follows, consistent with the baseline study:the value of K was 6, the original image size was 1024×512 pixels, and the loss function used was categorical cross-entropy, which estimates the

probability between the predicted voxels and the ground truth. The batch size was set to 4, AdaBound was used as the optimizer, the learning rate was 0.001, and early stopping was employed to prevent overfitting.

During testing, the Dice score was calculated for each patient in the test fold (4 patients per fold), and the mean value was taken across patients in each fold, considering only those patients with segmentation references. At least one of the IRF or SRF fluids is missing in some patients (both fluids were missing in patient 17, and one fluid is missing in patients 3, 9, 13, 16, 20, 21, 22, and 24). Therefore, during testing, for patients with at least one missing fluid in the B-scans, the Dice score for that class and that patient was excluded to avoid overestimation or underestimation. In cases where at least one fluid was missing in some B-scans but not all of them, the Dice score for those scans was set to zero. The Dice score was calculated per patient because mixing adjacent B-scans of the same patient with those of other patients could lead to overestimation.

2) **Duke DME**:
For the Duke DME, training and testing were conducted using annotations from Expert 2. Training was carried out on 55 B-scans, with no data augmentation applied. B-scans were used instead of entire volumes due to the anisotropic resolution of OCT volumes and the potential presence of motion artifacts across B-scans. To ensure fairness, the parameters and environmental settings were kept consistent for both the proposed model and the comparison models. Each fold consisted of B-scans from 5 patients, with patients 1-5 in the first fold and patients 6-10 in the second fold. To avoid bias, adjacent B-scans were not used across training, validation, and testing.
For all experiments, the parameters were set as follows, in line with the comparison models: the value of K was set to 2, and the B-scans were resized to 512×512 pixels. The loss function used was categorical cross-entropy. The batch size was set to 4. The cost function was optimized using AdaBound and backpropagation with the chain rule. The model was trained for 200 epochs, with early stopping employed to prevent overfitting.

In both datasets, the training and testing split was performed on a per-patient basis to prevent the mixing of B-scans from the same patient across both sets, thereby reducing the risk of overfitting. In the AROI dataset, the number of B-scans varies per patient, with the hold-out set consisting of B-scans from 4 patients (unseen patients at training), while B-scans from the remaining 20 patients were used for training. In the Duke DME dataset, a 50-50 split was applied, where B-scans from 5 unseen patients (55 B-scans) were used as the hold-out set, and the remaining 55 B-scans from another 5 patients were used for training.

The fluid class was absent in some B-scans for certain patients. Consequently, during testing, the Dice score calculation for the fluid class was excluded for B-scans that lacked fluid references for that patient, to avoid overestimation or underestimation.

The Friedman test was used to detect differences in performance between the algorithms evaluated on the segment classes. We computed the Friedman test statistic to check for significant differences and ranked the algorithms per segment class.

Deep_ResUNet++ was trained for 2 hours on an NVIDIA RTX A5000 GPU workstation. The code was implemented in Python, using the Keras library with the TensorFlow backend.

4.4.4 Results

The experimental results are presented for each dataset as follows:

1) **Annotated Retinal OCT Images (AROI)**:
Here, we report the performance measured by the Dice score for each segment class and method, including the Inter-observer, the baseline U-Net, the proposed Deep_ResUNet++, and other state-of-the-art architectures (nnUNet_RASPP, ResUNet, and ResUNet++) in this domain.
The segmentation performance, grouped by segment class, is illustrated in Fig. 4.5, and the corresponding Dice scores are presented in Table 4.1. Examples of the segmentation results are shown in Fig. 4.6, alongside the original input images and their annotations. We performed the Friedman test to determine whether the results were statistically significant and ranked the algorithms based on segment classes, obtaining a p-value of 0.0142 and a Friedman test statistic of 14.2381. The result, including algorithm rankings and scores, are presented in Table 4.3.
From these results, we observe the following:

1. The proposed model Deep_ResUNet++ achieved a mean Dice score of 0.98 outperforming the second best architecture by 0.02 (2%).

2. The proposed model Deep_ResUNet++ outperforms the baseline (U-Net), and current state-of-the-art models (nnUNet_RASPP, ResUNet, and ResUNet++) in every single class, achieving a Dice score above 0.90.

3. The IRF class was the most difficult to segment, with the Deep_ResUNet++ achieving a Dice score of 0.91, which is 11.5% higher than that achieved by the second-best model, ResUNet++, for that class.

4. An increase in performance is observed from the standard U-Net to more complex architectures, in the order of ResUNet, nnUNet_RASPP, ResUNet++, and Deep_ResUNet++.

5. It is also observed that the Deep_ResUNet++ obtained an overall mean Dice score of 0.98, which is 0.1(10%) higher than the human experts' annotation results of 0.88.

6. The Dice scores for the background classes and the layer classes (except RPE/BM) were consistently very high for all the models. This was expected, as the background classes occupy most of the image, and the two other layers, except RPE/BM, are made up of three or more thick retinal layers, whereas the RPE/BM consists of two thin retinal layers.

Table 4.1: Table of Dice Scores organized by segment classes (rows) and models (columns).

	Inter_Ob.	U-Net	ResUNet	ResUNet++	nnUNet_RASPP	**Deep_ResUNet++ (Proposed)**
Above_ILM	0.982	0.995	0.9991	0.9996	0.9991	**0.9998**
ILM	0.95	0.95	0.9859	0.9953	0.9892	**0.9973**
IPL_INL	0.948	0.923	0.9843	0.9947	0.9723	**0.9956**
RPE_BM	0.699	0.669	0.8907	0.9599	0.9212	**0.9640**
Under_BM	0.989	0.988	0.9993	0.9997	0.9991	**0.9998**
PED	0.860	0.638	0.9594	0.9846	0.9741	**0.9902**
SRF_SRHM	0.876	0.531	0.8805	0.9543	0.882	**0.9615**
IRF	0.735	0.48	0.7233	0.794	0.7757	**0.9098**
Mean	0.88	0.77	0.93	0.96	0.94	**0.98**

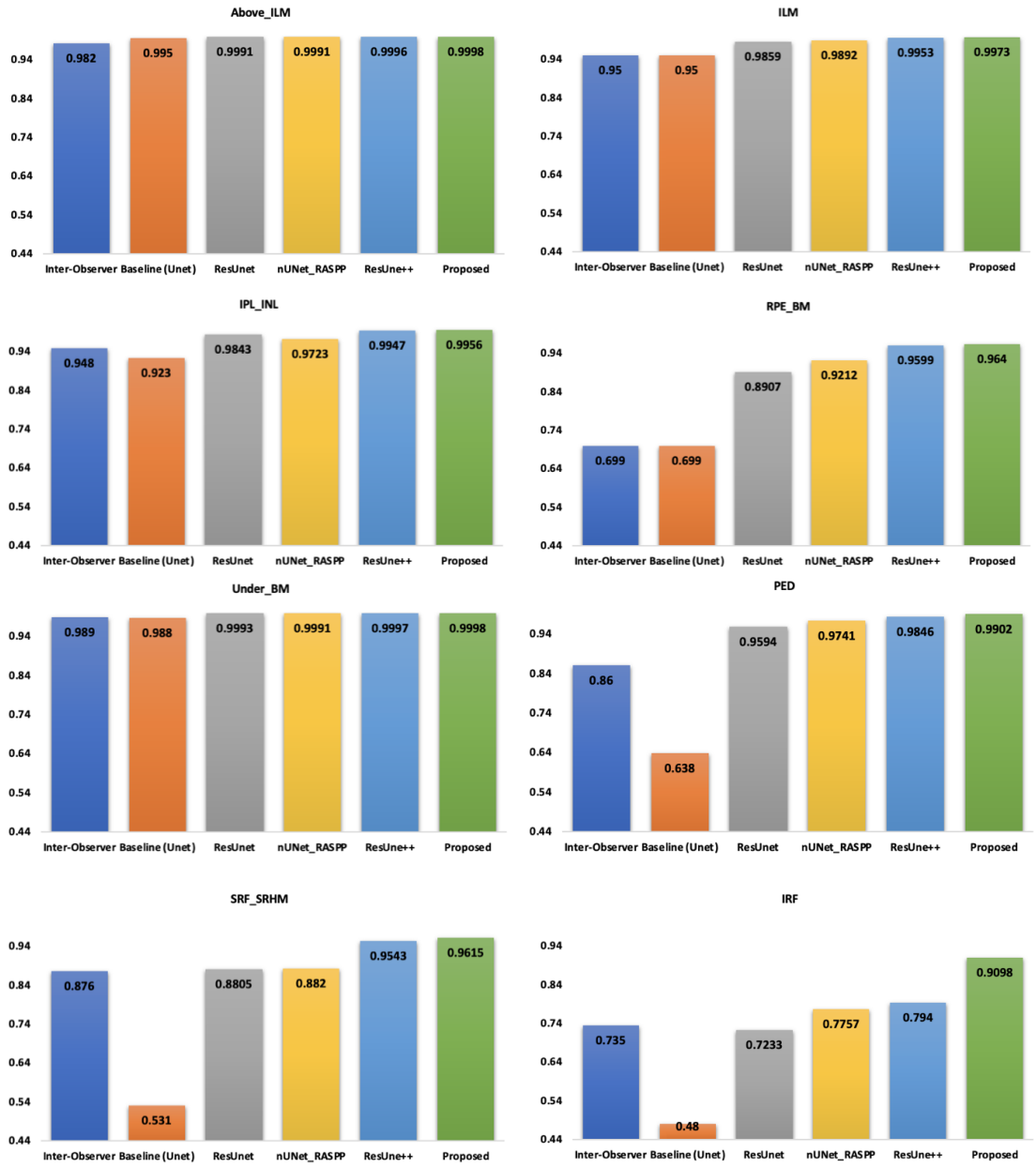

Figure 4.5: Performance comparison (measured by Dice scores) of the proposed Deep_ResUNet++ (Proposed) method, the baseline U-Net model, the Inter-Observer (by human experts), and other state-of-the-art models: UNet_ASPP, ResUNet, and ResUNet++ in this domain. The results are grouped by segment class.

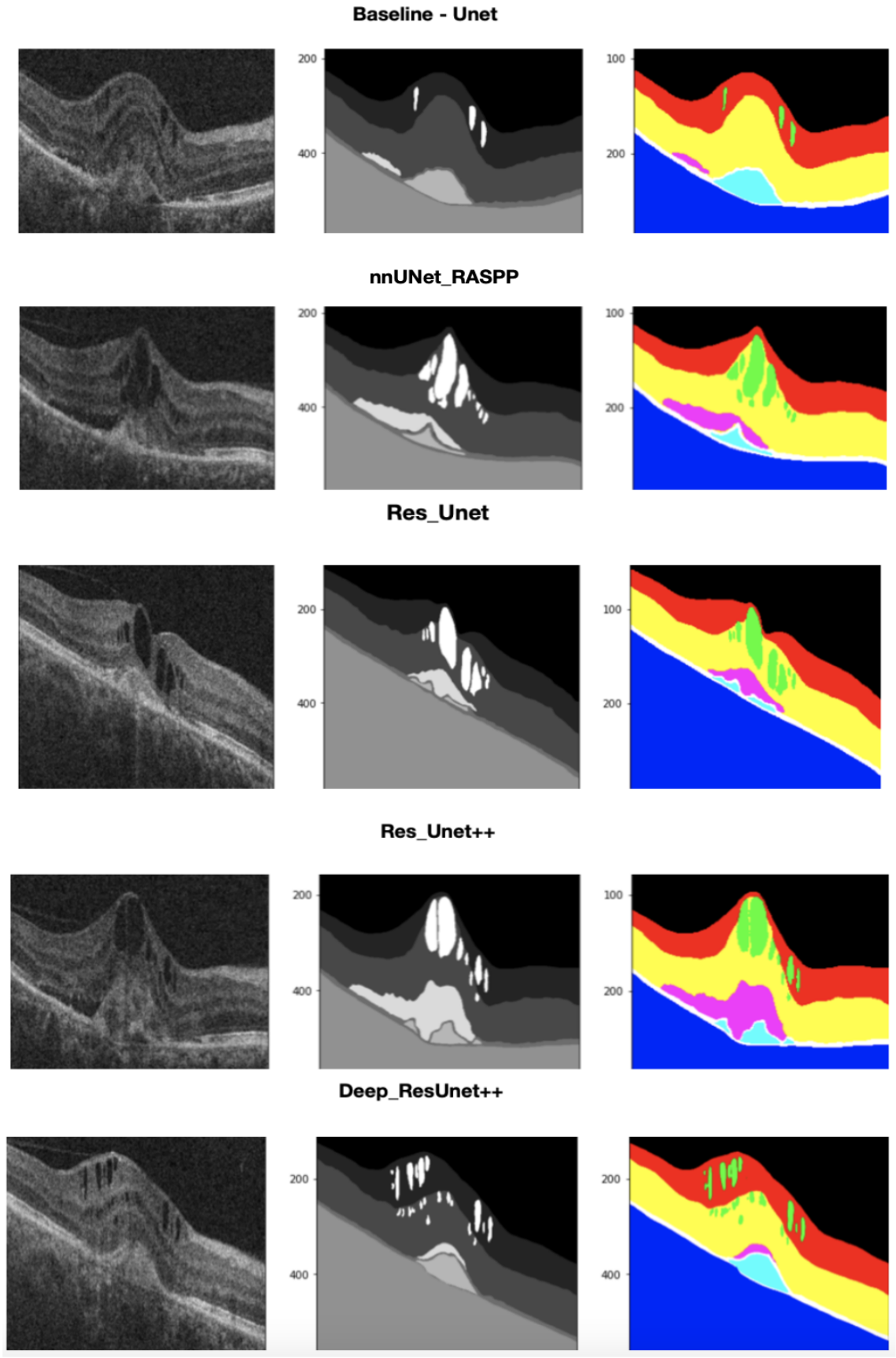

Figure 4.6: Examples of segmentation results, shown from left to right, include the inputs, annotations, and outputs for the Baseline U-Net, three state-of-the-art models, and the Deep_ResUNet++ (Proposed).

2) **Duke DME**:

Here, we present and analyze the segmentation class results measured by the Dice score for the Deep_ResUNet++(Proposed) on the Duke DME dataset. We compare these results to those of the comparison models (state-of-the-art models, ReLayNet, and the baseline U-Net), as well as to the human expert annotations (inter-observer) for this dataset.

The Dice scores are presented in Table 4.2, and the corresponding bar chart, grouped by segment classes, is shown in Figure 4.7. Examples of the visualization results, along with their annotations, are illustrated in Figure 4.8. A zoomed-in example of a visualization output from Deep_ResUNet++(Proposed) is provided in Figure 4.9. For the visualization results, orange arrows are used to highlight fine details in the annotated B-scans that were identified by the algorithms. Analysis of our results shows that:

1. The proposed model, Deep_ResUNet++, outperforms, the baseline (U-Net), and the current state-of-the-art model, ReLayNet, in every single class by a clear margin.

2. Deep_ResUNet++ achieved a Dice Score of 0.77, which is 0.19(19%) higher than the inter-observer Dice Score from human experts for the fluid class, which was the most challenging to segment.

3. Deep_ResUNet++ achieved a Dice Score of 0.90 or higher in 8 out of the 10 classes.

4. All the models achieved a perfect Dice Score of 1 for both background classes (the areas above and below the retina).

5. Deep_ResUNet++ achieved an overall mean Dice Score of 0.88, which is 0.8 higher than the 0.80 obtained from human experts' annotations (inter observer).

Table 4.2: Segmentation performance, measured by Dice Scores, organized by segment classes (rows) and models (columns).

	Inter_Obs.	U-Net	ResUNet++	nnUNet_RASPP	ReLayNet	**Deep_ResUNet++ (Proposed)**
Fluid	0.58	0.70	0.71	0.72	0.75	**0.77**
NFL	0.86	0.85	0.82	0.84	0.88	**0.90**
GCL_IPL	0.89	0.90	0.87	0.89	0.92	**0.93**
INL	0.77	0.77	0.76	0.78	0.82	**0.83**
OPL	0.72	0.74	0.73	0.76	0.80	**0.82**
ONL_ISM	0.87	0.88	0.87	0.89	0.91	**0.93**
ISE	0.85	0.86	0.86	0.88	0.92	**0.93**
OS_RPE	0.82	0.84	0.81	0.85	0.89	**0.91**
Mean	0.80	0.82	0.80	0.83	0.86	**0.88**

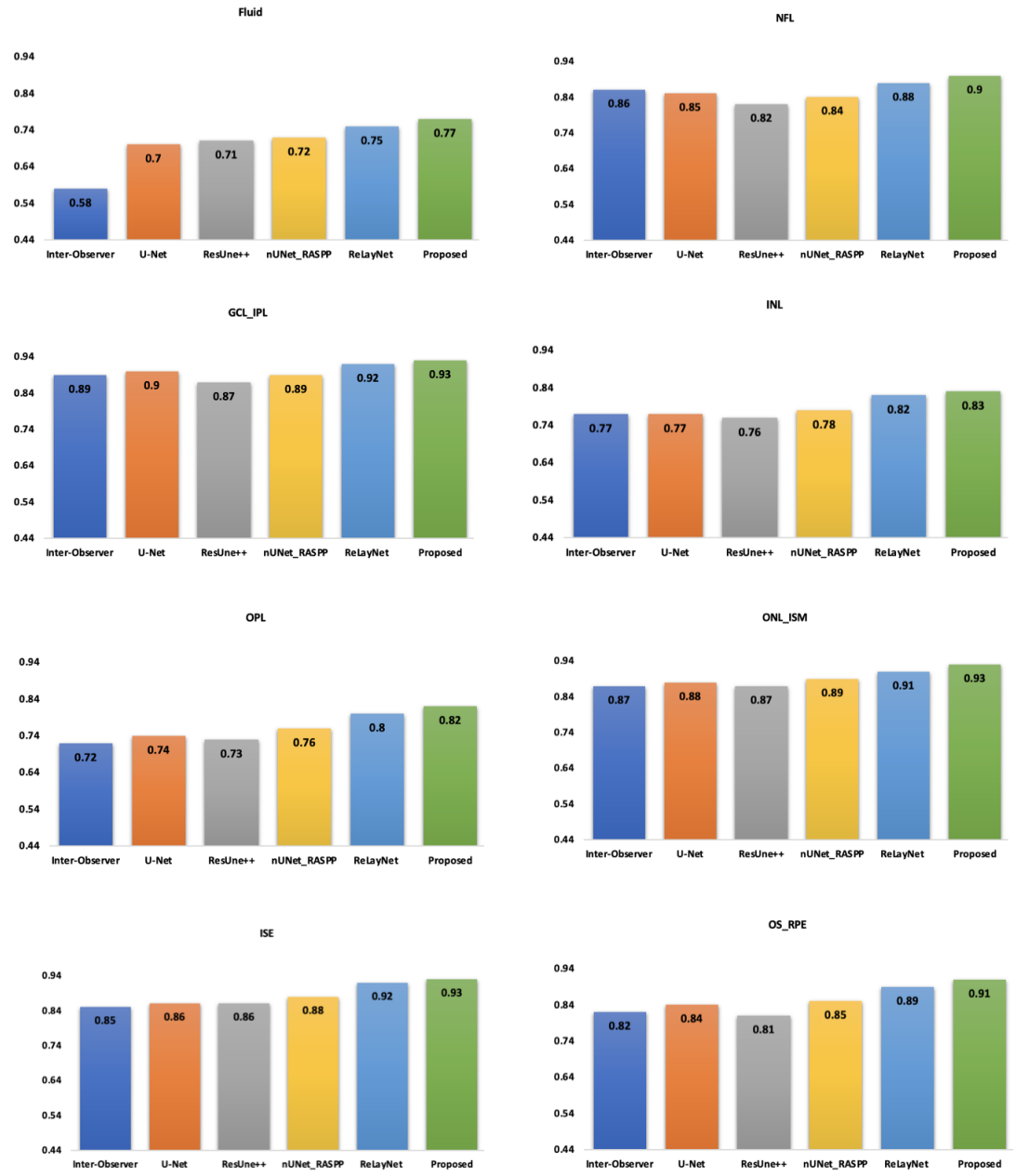

Figure 4.7: Bar chart comparison of Dice score performance, grouped by segment class, for inter-observer, U-Net, ResUNet, ResUNet++, ReLayNet, and the proposed Deep_ResUNet++ (Proposed) model.

Number of segment classes: 3
Number of algorithms : 6
Degrees of freedom: (5, 2)
Significance level (alpha): 0.05
p-value: 0.0142
Friedman statistic: 14.2381
Hypothesis: Alternative Hypothesis
Significant: There is a significant difference between at least two algorithms (p-value < 0.05).

Rank	Algorithm	Ranking Score
1	Deep_ResUNet++	1.00
2	ResUNet++	2.33
3	nnUNet_RASPP	2.67
4	ResUNet	4.00
5	Inter Ob.	5.33
6	U-Net	5.67

Table 4.3: The ranking (from best to worst) of the teams/algorithms based on the Dice Score (DS), using the Friedman test (a non-parametric test) indicates a significant difference between at least two of the algorithms, with a p-value of 0.0142 < 0.05 and a Friedman test statistic of 14.2381.

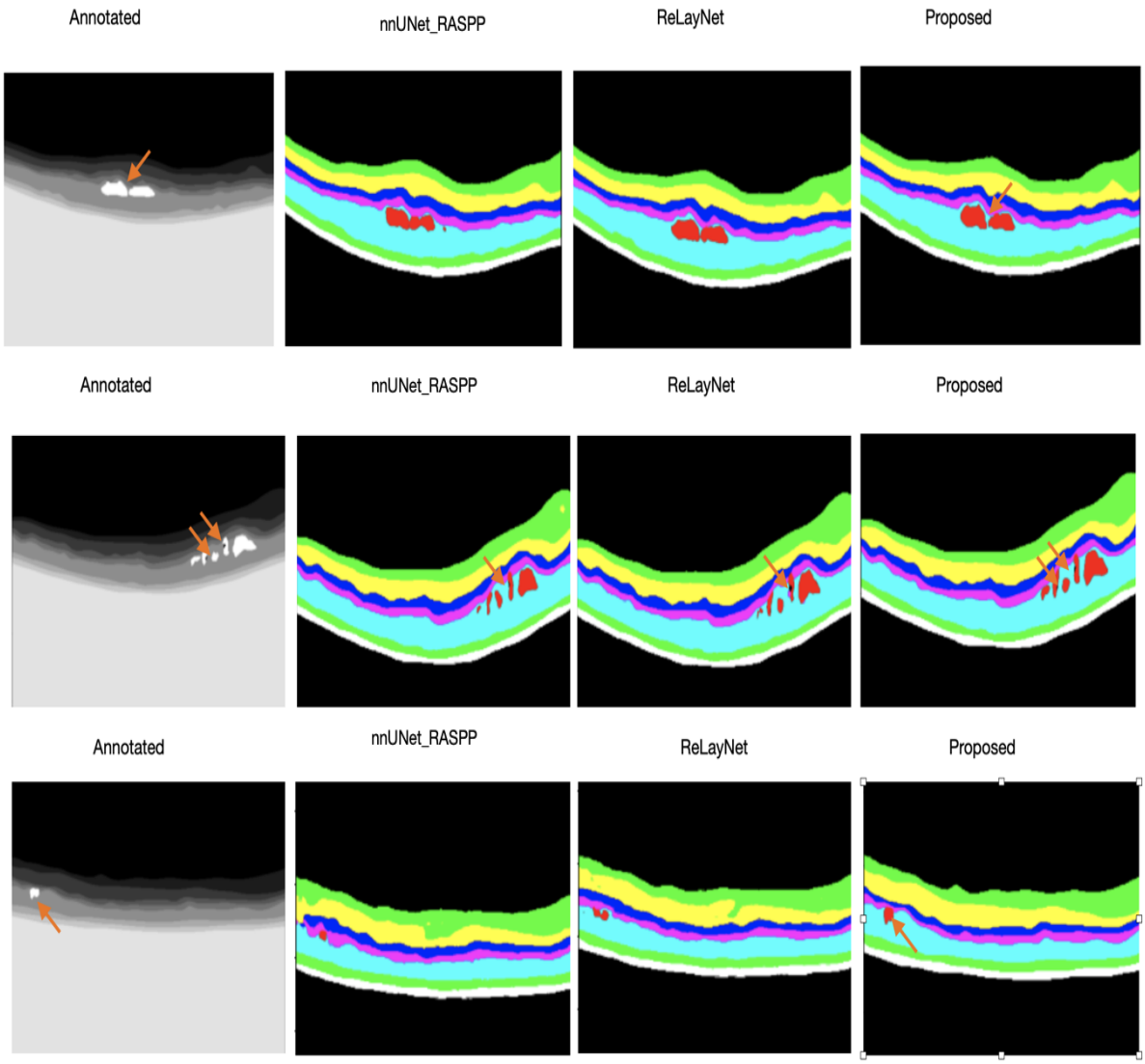

Figure 4.8: Examples to illustrate the visualisation output of the top three best performing algorithms: U-Net, ReLayNet and Deep_ResUNet++ (Proposed), in order of the inputs, annotations and outputs with orange arrows to demonstrate fine details picked up by the models.

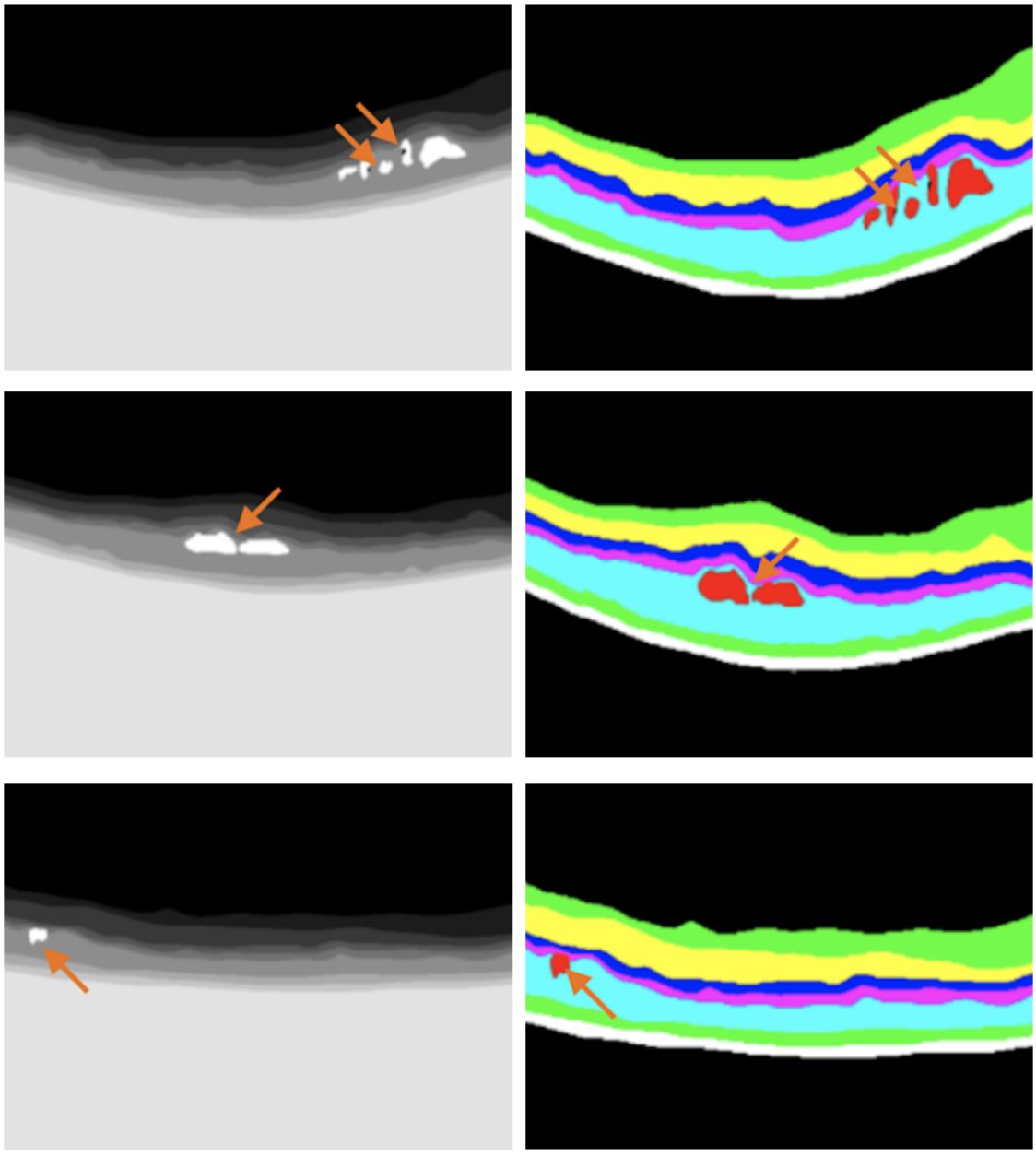

Figure 4.9: A zoom-in of the B-scan from Figure 4.8, highlighting the fine details identified by Deep_ResUNet++(Proposed) using orange arrows.

4.5 Summary

In this chapter, we investigated the simultaneous segmentation of retinal layers and diseases or fluid regions in retinal OCT images. To address this, we proposed a novel architecture, Deep_ResUNet++, by enhancing the ResUNet architecture. The enhancements involved: (i) dynamically adjusting the kernel size and network depth based on the input image size, and (ii) integrating multiple Atrous Spatial Pyramid Pooling (ASPP) blocks at the input, bridge, and output layers, along with Squeeze-and-Excitation (SE) blocks and a dense layer. The modifications allowed the network to effectively capture global contextual information at varying rates and across different locations, spanning multiple resolutions and depths. This approach enhanced the model's generalizability by: (i) effectively capturing structures of varying sizes within the images, (ii) adapting to diverse dataset characteristics, such as variations in resolution and noise, (iii) capturing both local and global context for a more comprehensive understanding of the data, and (iv) reducing the model's dependence on features from any single scale. As a result, Deep_ResUNet++ significantly improves segmentation accuracy and generalization performance. The algorithm was evaluated on two publicly available benchmark datasets, representing patients with two types of diseases: age-related macular degeneration (AMD) from the AROI dataset and diabetic macular edema (DME) from the Duke dataset. Experimental results demonstrate that Deep_ResUNet++ outperformed the baseline U-Net and other state-of-the-art methods on both datasets. The method presented in this chapter has practical applications for the structural analysis of OCT retinal images and for monitoring the progression of eye diseases such as AMD and DME.

Chapter 5

Enhancing Medical Image Segmentation Through Knowledge Transfer with Domain-Specific Adapters Across Diverse Data Sources

In Chapters 3 and 4, we explored the potential of incorporating Atrous Spatial Pyramid Pooling (ASPP) blocks at various positions in the Convolutional Neural Network (CNN) backbone to capture global contextual information and improve the model's generalization performance, while overlooking the combined potential of other available annotated datasets. One approach to improve the generalization performance on unseen datasets is to build a single, diverse model by integrating data from multiple sources, organs, modalities, and disease types. While numerous small annotated medical image datasets are publicly available across various modalities, organs, and diseases, naively combining data from diverse sources can negatively impact the model performance due to the transfer of negative knowledge from one dataset to another. In this chapter, we explore the synergistic potential of combining data from multiple diverse sources, modalities, organs, and disease types to build a single, generalizable model. To mitigate negative knowledge transfer, we employ domain-specific knowledge transfer adapters. In deep learning, the two predominant approaches for medical image segmentation are Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). This chapter introduces two novel methods that leverage knowledge transfer and domain-specific adapters. The first algorithm utilizes a CNN, while the second combines CNNs with ViTs in a hybrid approach to demonstrate the effectiveness of domain-specific adapters in multi-source medical image segmentation and detection. These methods are: 1) MMIS-Net (MultiModal Medical Image Segmentation Network), which incorporates Similarity Fusion blocks to use supervision and pixel-wise similarity for feature map fusion for knowledge transfer into a CNN backbone. To address inconsistent class definitions and label contradictions, we developed a one-hot label space to handle classes absent in one dataset but annotated in another. This approach preserves distinct annotation protocols for the same target structure during training. 2) CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-

Specific Batch Normalization), which combines CNNs for feature extraction, Vision Transformers for capturing long-range dependencies, and to address negative knowledge transfer, within the network we integrated domain-specific adapters to capture and share domain specific information across all domains, hence reducing negative knowledge transfer between domains. Both approaches were evaluated on two dataset groups. The first group comprises of 10 benchmark datasets covering 19 organs across 2 modalities, and the second group is the HECKTOR 2022 dataset collected from 9 medical centers worldwide. Experimental results show that: (i) On the RETOUCH Grand Challenge hidden test set MMIS-Net outperformed state-of-the-arts (SOTA) architectures and large foundation models for medical image segmentation, achieving a mean Dice score (DS) of 0.83 and an absolute volume difference (AVD) of 0.035 for retinal fluid segmentation, as well as a perfect area under the curve (AUC) of 1.0 for fluid detection. (ii) CVD_Net achieved a mean Dice score of 0.77492, comparable to state-of-the-art performance, on the HECKTOR 2022 hidden dataset, which includes data from two new medical centers not seen during training. Both models also demonstrated high generalization performance when tested on independent data from new sources not seen during training.

Some of the work presented in this chapter is currently under review for a journal publication under the title " MMIS-Net for Retinal Fluid Segmentation and Detection" and a conference paper published in [137]. The manuscripts are lead-authored by the author of this thesis, who made substantial contributions to the conception, data collection, processing, and writing of the manuscripts, as well as sole contributions to the implementation and result analysis.

5.1 Introduction

Image segmentation is a widely studied problem in the deep learning community and is paramount in medical image analysis, diagnostics, and monitoring the progression of pathogens/diseases. Medical image segmentation tasks involve diverse modalities such as Optical Coherence Tomography (OCT), Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Ultrasound, X-ray, and many more, incorporating various anatomical structures such as the retina, brain, neck, fetal tissues, chest, abdomen, cells, and more. Several small datasets with their corresponding annotations/labels from different modalities and anatomic regions are available in the public domain. This availability has sparked the development of numerous deep learning algorithms for lesion segmentation in medical imaging. However, most of these algorithms are typically trained on a single modality for a specific anatomic structure or problem, leading to challenges in generalization to new, unseen datasets like in real-world scenarios. One of the main causes of this issue is the high variability in image quality stemming from different modalities, collected across various medical centers using machines from different manufacturers and annotated by radiologists with varying levels of experience. One approach to circumventing this problem is to increase the diversity of the training set by combining images from various modalities, representing different anatomic structures, and collected across different medical centers using devices from various vendors. The two most common approaches for medical image segmentation are Convolutional Neural Networks (CNN) and, more recently, Vision Transformers (ViT). According to [51], CNN demonstrates superior performance on smaller datasets, while ViT tends to outperform CNN on larger datasets due to their intrinsic ability to model long-range dependencies, although they require more data. We aim to construct a universal model that generalizes across multiple data sources by integrating datasets from various domains. However, naively combining datasets from different sources can improve performance on one dataset while potentially reducing it on another, a phenomenon known as Negative Knowledge Transfer (NKT) [3], [213]. To mitigate the risks associated with NKT and to address variations in image quality, thereby enhancing generalization across diverse data sources. We have incorporated Domain-Specific Adapters (DSA) [29] into the network's architecture. In this chapter our main contributions are as follows:

1) We introduce MMIS-Net, a novel algorithm designed to train a single model to segment multiple lesions from various body structures across diverse image modalities simultaneously. MMIS-Net incorporates similarity fusion blocks into its architecture, utilizing supervision and pixel-wise selection knowledge for feature map fusion. This approach reduces irrelevant and noisy signals in the output.

2) We efficiently created a one-hot label space to address the inconsistent class definitions and label contradiction problem, covering diverse modalities and body regions in a multiclass segmentation problem. This strategy effectively manages classes that are absent in one dataset but annotated in another during training. Also, it retains different annotation protocol characteristics for the same target structure and allows for overlapping target structures with different levels of detail, such as liver, liver vessels, and liver tumors.

3) We introduce CVD_Net, a novel architecture combining CNNs for feature extraction, ViTs for capturing long-range dependencies, and domain-specific adapters

to capture and share domain specific information across all domain. This reduces negative knowledge transfer between domains while enhancing the model's generalization ability.

The rest of this chapter is organized as follows. The proposed methods are detailed in Sections 5.2 and 5.3. Section 5.4 discusses the datasets, experiments, results, and visualizations. Finally, the summary is provided in Section 5.5.

5.2 MMIS-Net Method

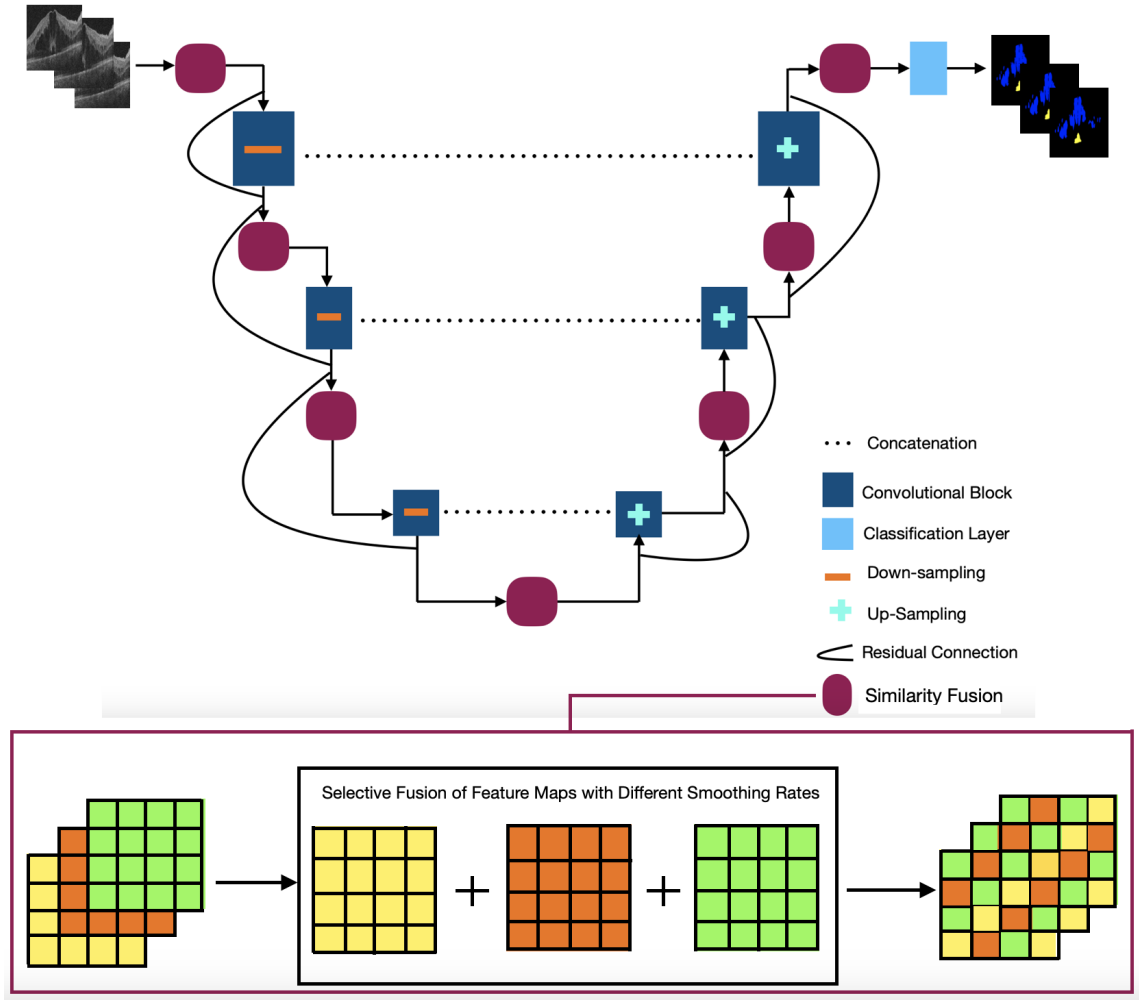

Figure 5.1: A high-level illustration of the MMIS-Net architecture demonstrating the contracting and expanding paths, residual connections, and the similarity fusion blocks. Further details of the fusion block, illustrating the feature map fusion using supervision and pixel-wise similarity selection of images at different smoothing scales, is shown at the bottom.

In a collection of multiple datasets, each pixel is assigned to a segmentation class with corresponding label pairs. For each dataset in the collection, every pixel within the raw image is associated with a segmentation class, which is then mapped to a label value in the annotated dataset for that specific dataset.

We combined all the label images into a single one-hot label space for all the datasets and each class is assigned a unique label value as demonstrated in Table 5.3. Combining partially annotated datasets presents its own challenges, and here are some: 1) Label Index Inconsistency: The same organ can be labeled with different indexes in different datasets. 2) Background Inconsistency: An organ is marked as background in one dataset but as foreground in another. For example, in the Pancreas-CT dataset [142], the pancreas is marked as foreground, but it is marked as background in the MSD Spleen dataset [129]. 3) Absent of Organ Labels: The same organ is labeled in one dataset but absent in another dataset that also contains the organ. For example, in the MSD Liver dataset, both the liver and liver tumor are segmented. In contrast, in the MSD Hepatic Vessels dataset, the labeled targets are the vessels and tumors within the liver, but not the liver itself. 4)Organ overlapping. There is overlap between sub-structures and organs. For example, in one dataset, the Hepatic Vessel, a sub-structure of the Liver, is segmented separately from the Liver, while in another dataset, both the Hepatic Vessel and the Liver are annotated as separate organs. A similar case occurs with the Kidney and Kidney Tumor. Various methods, such as [110], have tried to address these challenges by combining labels with text embedding and adopting a masked back-propagation mechanism. In this work, we use labels only and enhance the network architecture to effectively address the partially labeled class problem, where certain classes are labeled in one dataset but not in another for the same organ during training. Our strategy also mitigates the issue of overlapping target structures, such as the liver, liver vessels, and liver tumors, by preserving the unique characteristics of different annotation protocols for the same target structure. This approach ensures that, in cases where one or more classes from different datasets refer to the same structure, the network treats them as distinct. This accounts for the unknown and potentially variable annotation protocols and labeling characteristics across datasets. Consequently, the network must be able to predict multiple classes for a single voxel/pixel to accommodate these inconsistent class definitions. To address the label contradiction problem, unlike the commonly used Softmax, we employed the Sigmoid activation function to separate the outputs for each class. During training, each class was assigned an independent segmentation head with parameters that share a common backbone within the network. This enhancement enables the architecture to segment overlapping classes while preserving all label properties from each dataset by assigning multiple segmentation classes to a single pixel. At the classification layer, this adjustment can be thought of as a binary segmentation problem.

The MMIS-Net (MultiModal Medical Image Segmentation Network) is composed of five main components: a contracting path (the encoder), an expansion path (the decoder), the similarity fusion block, residual connections, and a class-adaptive loss function.

The Contracting Path

The contracting path is used to capture contextual information and as we go down the contracting path the image is halved after every convolutional block. Each block consists of two 3x3 convolutions followed by a ReLU (Rectified Linear Unit) activation function and next is followed by a 2x2 max-pooling, which reduces the feature map by half.

The Expanding Path

The expanding path is used for pixel localization. As we go up the expanding path, the feature map is doubled after every convolutional block by concatenating the feature map of the expanding path with its corresponding map in the contracting path. Each block in the expanding path is composed of a 2x2 transpose convolution, followed by a concatenation, two 3x3 convolutions, and a ReLU activation function.

Similarity Fusion Blocks

The Similarity Fusion is a technique aimed at capturing cross-dimensional dependencies in feature maps and handling datasets with inconsistent labels. This approach effectively models complex relationships across input dimensions, facilitating improved representation learning and feature extraction by exploiting correlations between spatial, temporal, or channel-wise relationships. Unlike the standard fusion module [80], which achieves feature fusion through pixel-wise summation or channel-wise concatenation, the similarity fusion block uses supervision and selection similarity knowledge to reduce irrelevant and noisy signals in the output. This is crucial for capturing the synergistic potential of diverse datasets from multiple modalities, encompassing different organs with various diseases, and for mitigating negative knowledge transfer during training. Given an input image, we enhance its quality and remove noise by applying a Gaussian filter [85] at various smoothing rates using different sigma values, producing three new images. To further reduce the noise, we use the Euclidean distance similarity measure [191] at the pixel level to calculate the similarity. Pixels from the same position on all three images are grouped together. Each group contains three pixels, one from each of the three different feature maps. The pixel similarity is measured at the group level. Within each group, the pixel that is most similar to the other two is chosen, while the other two are excluded. The similarity is measured by finding the pixel with the shortest distance to the other two. The similarity fusion block is integrated into the network's architecture before and after every convolutional block in both the contracting and expanding paths. It is also used in the bridge layer. A high level diagram to demonstrate the similarity block is shown at the bottom of Figure 5.1 and a snippet of the similarity fusion pseudocode is shown in Listing 1.

Algorithm 1 Snippet of the Similarity Fusion Pseudocode

for each fusion map **do**
    Generate three fusion maps at different smoothing scales
    for each pixel **do**
        for each position along the Z-axis **do**
            Compute the similarity between pixels using the distance matrix
            Select the two pixels with the shortest(minimum) distance
            Fuse selected pixels across the Z-axis using Euclidean distance
            Compute Euclidean distances:
                $d_1 = \sqrt{(O_1 - O_2)^2 + (G_1 - G_2)^2 + (Y_1 - Y_2)^2}$
                $d_2 = \sqrt{(O_1 - O_3)^2 + (G_1 - G_3)^2 + (Y_1 - Y_3)^2}$
                $d_3 = \sqrt{(O_2 - O_3)^2 + (G_2 - G_3)^2 + (Y_2 - Y_3)^2}$
            Select minimum distance:
                $d_{\min} = \min(d_1, d_2, d_3)$
        end for
    end for
end for

The Residual Connection

Residual connection [68] is a skip connection that enables the network to learn residual mappings instead of directly fitting the desired underlying mapping. Traditional deep networks aim to approximate the underlying mapping $H(x)$ using stacked layers. However, during training, it can be challenging for deeper networks to learn these mappings effectively. Residual learning introduces the concept of learning residual functions, denoted as $F(x) = H(x) - x$, where $H(x)$ is the desired mapping and x is the input to a certain layer. The residual connection is incorporated into the network's architecture at every level in both the contracting and expanding paths to mitigate the problem of vanishing gradients.

The Class-adaptive Loss Function

The loss function used is a combination of cross-entropy and Dice loss. We employed binary cross-entropy loss and a modified Dice loss.The regular dice loss is calculated individually for each image in a batch, whereas we jointly calculate the dice loss for all images in the input batch. This approach helps regularize the loss when only a few voxels of one class appear in one image, while a larger area is present in another image of the same batch. Consequently, inaccurate predictions of a few pixels in one image have a limited impact on the overall loss.

Between the contracting and expanding paths is a bridge layer composed of a similarity fusion block to ensure a smooth transition from one path to the other. At the end of the expanding path is a classification layer to classify each pixel as belonging to the background or one of the segmented classes.

Hyperparameter Settings

The encoder path consists of convolutional blocks, each containing a $3 \times 3 \times 3$ convolutional layer, Batch Normalization (BN), ReLU activation, max pooling with a $2 \times 2 \times 2$ kernel, and zero padding with a stride of 2. The decoder path mirrors

the encoder's structure but replaces max pooling with upsampling and includes concatenation with corresponding encoder features for precise reconstruction. The network depth was set to 4, corresponding to an input size of $512 \times 512 \times 432$ pixels. The Similarity Fusion Blocks utilized Euclidean distance for pixel selection. The loss function was a combination of cross-entropy and Dice loss, the learning rate was set to 0.1, the batch size was 4, the optimizer used was Adam, and the maximum training epoch was set to 1000, with early stopping used to avoid overfitting.

5.3 CVD_Net Method

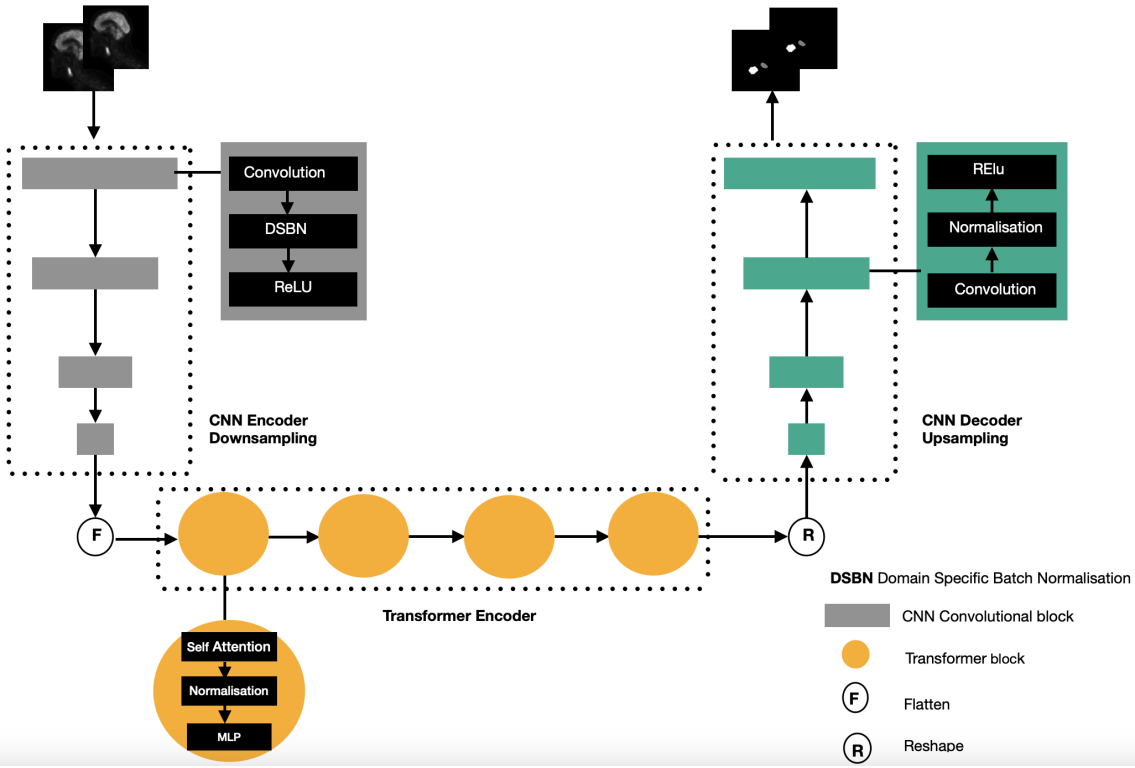

Figure 5.2: A high-level illustration of the CVD_Net architecture. The convolutional blocks at the CNN encoder for feature map extraction are shown in gray, those at the CNN decoder for upsampling in green, and the Transformer blocks to capture long-range dependencies at the encoder in yellow. F stands for flattening the maps before feeding into the Transformer encoder, and R stands for reshaping the maps before feeding into the CNN decoder.

CVD_Net is composed of four main components: a CNN encoder, a domain-specific batch normalization, a Transformer encoder, and a CNN decoder as demonstrated in Figure 5.2. Details of these components are as follows.

CNN Encoder

The CNN Encoder is used to extract features from the input images and it is composed of three convolutional blocks in series with residual connections. Each of the block is followed by a batch normalization and and Rectified Linear Unit (ReLU) activation. Given a raw image $X \in \mathbb{R}^{H \times W \times D}$ whose spatial resolution is H $\times$ W and

the depth (number of slices) D. The feature maps produced by the CNN Encoder (F_CNN) can be formally expressed as :

$$\{f\}L_{l=1} = F^{l}_{CNN}(x;\Theta) \in \mathbb{R}^{C\times D\times \frac{2}{l}\times H\times \frac{2}{l+1}\times W\times \frac{2}{l+1}} \tag{5.1}$$

where $\{f\}L_{l=1}$ is the feature map produced by the CNN Encoder, x is the input image, L indicates the number of feature levels, Θ denotes the parameters of the CNN encoder, and C denotes the number of channels.

Domain-specific Batch Normalization (DSBN)

The DSBN [29] is the batch normalization technique used at every convolutional block within the CNN encoder to capture domain-specific information. The DSBN consists of several batch normalization layers, each reserved for a specific domain and a shared parameter backbone that retains common shared parameters to learn general features applicable across all domains. By leveraging the generalizable features learned by the shared backbone, and effectively mitigating negative knowledge transfer, DSBN ensures that the model is not biased toward dominant domains and enhances the model's ability to generalize across diverse domains, thereby improving its overall generalization capability.

Transfomer Encoder

The Transformer encoder is used to capture long-range dependencies from the extracted features. It is composed of an input-to-sequence layer and stacked deformable Transformer layers. The extracted features from the CNN encoder are flattened into a 1D vector before being fed into the Transformer encoder. Due to this, they lose some spatial information. To mitigate this problem, we employ sine and cosine functions with different frequencies [187] to compute the positional coordinates of each dimension. The Transformer encoder consists of transformer blocks stacked in series. Each transformer block employs the self-attention mechanism [219] to capture long-range dependencies by computing the weighted sum of the input data based on the similarity between the input features. The self-attention mechanism generates a trainable associative memory with a query (Q) and a pair of key (K)-value (V) pairs to produce an output by linearly transforming the input. This is represented as follows:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5.2}$$

where $\sqrt{d}$ is a scaling factor based on the depth of the network. The output is normalized and fed through a feed-forward multi-layer perceptron (MLP). Skip connections are employed to avoid the vanishing gradient problem.

CNN Decoder

The output of the transformer encoder is fed to a CNN decoder. The CNN decoder progressively upsamples the feature map through a series of convolutional blocks consisting of a convolutional layer, normalization layer, and a ReLU activation using residual connections. At the end of the decoder path is a classification layer for pixel classification. Assuming there are k classes, including the background, the classification layer predicts k semantic masks. $\hat{S}_l \in \mathbb{R}^{h\times w\times k}$ simultaneously,

corresponding to each semantic label as demonstrated on Eqn (5.3)

$$\hat{S} = argmax(Softmax(\hat{S}_l, d = -1), d = -1) \tag{5.3}$$

Where $d = -1$ indicates the Softmax and argmax operations performed across the last dimension (the channel dimension). The loss functions is the sum of the cross entropy and Dice loss which is express as follows :

$$L = \lambda_1 \text{CE}(\hat{S}, D(S)) + \lambda_2 \text{Dice}(\hat{S}, D(S)) \tag{5.4}$$

Where CE and Dice represents cross entropy loss and Dice loss, respectively. D denotes as the downsample operation. λ_1 and λ_2 represent the loss weights.

CVD_Net was evaluated on the first task of the second edition of the HEad and neCK TumOR (HECKTOR) challenge [7]. This task involves the automatic segmentation of Head and Neck primary Gross Tumor Volume (GTVt) in PET/CT images. The offline version of the challenge attracted 103 teams, resulting in 448 submissions. Here, we review some of the methods used by the top teams in the first task of this challenge. Team Pengy secured the first position, achieving a mean Dice score of 0.778. They utilized nnUNet [84], a self-configuring pipeline for medical image segmentation. SJTU [6] ranked second with a mean Dice score of 0.7733. Their method employed ResUNet [45] as a backbone, comprising three parts: the first part for extracting the region of interest (ROI), the second part for training a model based on the ROI, and the third part for refining the trained model. HiLab [119] presented an ensemble of five deep learning methods and an attention mechanism, achieving a mean Dice score of 0.773. BCIOqurit [208] extended nnUNet [84] by incorporating squeeze and excitation normalization [82] into the algorithm backbone, achieving a mean Dice score of 0.7709. Another nnUNet based method was presented by team Aarhus Oslo [153], achieving a mean Dice score of 0.779. Team Aarhus Oslo obtained the best DS but also had a high rate of missing predictions on one or multiple patients, hence it was ranked fifth by the organizers. The Fuller MDA [135] introduced an ensemble of 3D residual U-Nets trained on a 10-fold cross-validation and majority voting, obtaining a mean dice score of 0.7702.

Hyperparameter Settings

The CNN encoder path consists of 4 convolutional blocks, each containing a $3 \times 3 \times 3$ convolutional layer, Batch Normalization (BN), ReLU activation, max pooling with a $2 \times 2 \times 2$ kernel, and zero padding with a stride of 2. The extracted feature maps are then flattened into a 1D vector and passed into the Vision Transformer (ViT) encoder. The output is then reshaped back into 3D convolutions before being fed into the CNN decoder. The CNN decoder path mirrors the CNN encoder's structure but replaces max pooling with upsampling and includes concatenation with corresponding encoder features to enhance spatial information reconstruction. The DSBN consists of self-attention layers, normalization layers, and a Multi-Layer Perceptron (MLP). The network depth was set to 4, corresponding to an input size of $512 \times 512 \times 128$ pixels. The loss function was a combination of cross-entropy and Dice loss, the learning rate was set to 0.1, with a batch size of 4, and the optimizer used was Adam. The maximum training epoch was set to 1000, with early stopping.

5.4 Experiments

5.4.1 Dataset

In this chapter, two groups of datasets were used. The first group is a multi-organ dataset which comprises of 10 benchmark datasets covering 19 organs across 2 modalities, while the second group is the HECKTOR 2022 benchmark dataset, collected from 9 medical centers around the world.

Multi-organ

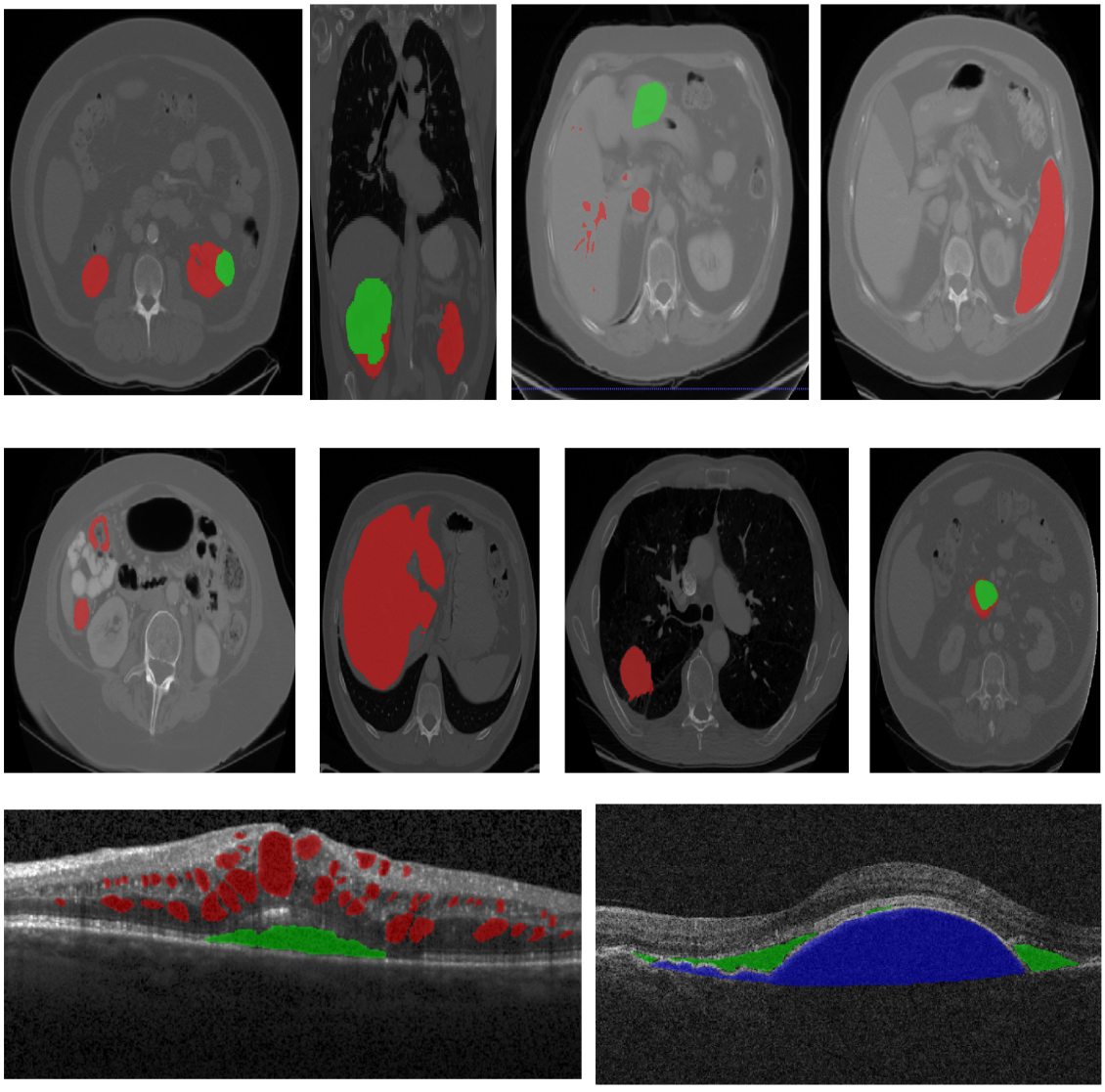

Figure 5.3: An illustration of B-Scans from different datasets of the Multi-organ dataset, showcasing various organs, modalities, and diseases, highlighting the high diversity of the datasets.

The Multi-organ dataset consist of a total of 10 datasets originating from the Medical Segmentation Decathlon (MSD) [129], Pelvis [177], Pancreas CT [142], KiTS19 [98], and RETOUCH [21], datasets. The datasets were annotated for 19

anatomic structures, consisting of 1337 volumes across 2 modalities: computed tomography (CT) and optical coherence tomography (OCT). These datasets cover 19 segmentation tasks and one detection task. The MSD datasets used are as follows: **Liver**: This dataset consists of 201 contrast-enhanced CT images from patients with primary cancers and metastatic liver disease. The segmented regions of interest are the liver and tumors inside. It was acquired at the IRCAD Hopitaux Universitaires, Strasbourg, France. **Pancreas**: This dataset consists of 421 CT scans of of patients undergoing resection of pancreatic masses. The segmented regions of interest are the pancreatic parenchyma and pancreatic mass (cyst or tumor). It was acquired at the Memorial Sloan Kettering Cancer Center, New York, USA. **Hepatic Vessels**: This dataset consists of 443 CT scans of patients with a variety of primary and metastatic liver tumors. The segmented regions of interest are the vessels and tumors within the liver. It was acquired at the Memorial Sloan Kettering Cancer Center, New York, US. **Lung**: This dataset consists of 96 CT scans of patients with non-small cell lung cancer, and the segmented region of interest is the lung tumors. It was collected from the Cancer Imaging Archive [24]. **Spleen**: This dataset consists of 61 CT scans of patients undergoing chemotherapy treatment for liver metastases, and the segmented region of interest is the spleen. It was acquired at the Memorial Sloan Kettering Cancer Center, New York, USA. **Colon**: This dataset consists of 190 CT scans of patients undergoing resection of primary colon cancer, and the segmented region of interest is the primary colon cancer. It was acquired at the Memorial Sloan Kettering Cancer Center, New York, USA. **KiTS19** [98]: This dataset consists of 300 CT scans. The segmented regions of interest are the kidneys and kidney tumors. They were acquired at the University of Minnesota Medical Center, USA. **Pelvis** [177]: This dataset consists of 50 CT scans, and the segmented regions of interest are the uterus, bladder, rectum, and bowel. The dataset was acquired from the Vanderbilt University Medical Center (VUMC), USA, and the Erasmus Medical Center (EMC) Cancer Institute in Rotterdam, the Netherlands. **Pancreas CT** [142]: This dataset consists of 82 CT scans, and the segmented region of interest is the pancreas. The dataset was acquired from the National Institutes of Health [142]. **RETOUCH** [21]: This dataset consists of 112 retinal optical coherence tomography (OCT) scans of patients with early age-related macular degeneration (AMD) and diabetic macular edema (DME), collected from three device vendors: Cirrus, Spectralis, and Topcon. For a fair comparison, the training set consisting of 70 scans is available to the public, and the testing set consisting of 42 hidden scans is held by the organizers. Submission and evaluation of predictions on the testing dataset are arranged privately with the organizers, and the results are sent to the participants. The dataset was segmented for three regions of interest: intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelium detachments (PED). The dataset was acquired from the Medical University of Vienna (MUV) in Austria, Erasmus University Medical Centre (ERASMUS), and Radboud University Medical Centre (RUNMC) in the Netherlands. Examples of the datasets are shown in Figure 5.3, and further details about the datasets' composition are provided in Table 5.1.

Datasets	Modality	Labels	Training	Shape	Spacing [mm]
Liver [129]	CT	Liver, L. Tumor	131	432x512x512	(1, 0.77, 0.77)
Lung [129]	CT	Lung nodules	63	252x512x512	(1.24, 0.79, 0.79)
Pancreas [129]	CT	Pancreas, P. Tumor	281	93x512x512	(2.5, 0.80, 0.80)
H. Vessels [129]	CT	H. vessels, H. Tumor	303	49x512x512	(5, 0.80, 0.80)
Spleen [129]	CT	Spleen	41	90x512x512	(5, 0.79, 0.79)
Colon [129]	CT	Colon cancer	126	95x512x512	(5, 0.78, 0.78)
Pelvis [177]	CT	Ut, Bl, Rec, Bow	30	180x512x512	(2.5, 0.98, 0.98)
Pancreas CT [142]	CT	Pancreas	82	217x512x512	(1, 0.86, 0.86)
KiTS19 [98]	CT	Kidney, K.Tumor	210	107x512x512	(3, 0.78, 0.78)
RETOUCH [21]	OCT	IRF, SRF, PED	70	128 x512x512	(0.01, 0.01, 0.05)
Total			**1337**		

Table 5.1: Summary table of the datasets used, showing the modalities, anatomic structures, number of training cases, median shapes, and image spacings. The abbreviations used in this table are L. Tumor, Liver Tumor; P. Tumor , Pancreas Tumor; H. Vessels, Hepatic Vessels; H. Tumor, Hepatic Tumor; Ut, Uterus; Bl, Bladder; Rec, Rectum; and Bow, Bowel.

HECKTOR 2022

Head and Neck (H&N) cancer is one of the most common worldwide and the fifth leading cause of death globally [145], accounting for 4% of all cancer deaths in the USA [173]. The head and neck squamous cell carcinoma (HNSCC) are the most common form of H&N cancers, typically originating in the squamous cells lining the mucosal surfaces of the mouth, throat, and voice box. Although head and neck cancers can also develop in the salivary glands, sinuses, or muscles and nerves in the head and neck, these types of cancer are much less prevalent than squamous cell carcinomas [41], [175]. Effective treatment plans for H&N cancers exist in the form of surgery, radiation therapy, chemotherapy, targeted therapy, immunotherapy, or a combination of these treatments. However, the effectiveness of these treatments depends on frequent monitoring and early detection of the disease.

The algorithms were evaluated on the HEad and neCK TumOR (HECKTOR) 2022 challenge benchmark dataset [7]. The dataset consists of 883 cases of PET/CT images collected from 9 medical centers from 4 different countries using 12 different medical devices. The dataset is split into 524 training cases from 7 different centers and 359 hidden test cases from 3 different centers, 2 of which are new centers not included in the training sets. The datasets were annotated by human experts for three classes: 0 for background, 1 for primary tumors (GTVp), and 2 for Gross Tumor Volumes (GTVn). A summary of the dataset is shown in Table 5.2. An illustration depicting the high variability in image quality among images sourced from the seven medical centers in the training set is shown in Figure 5.4.

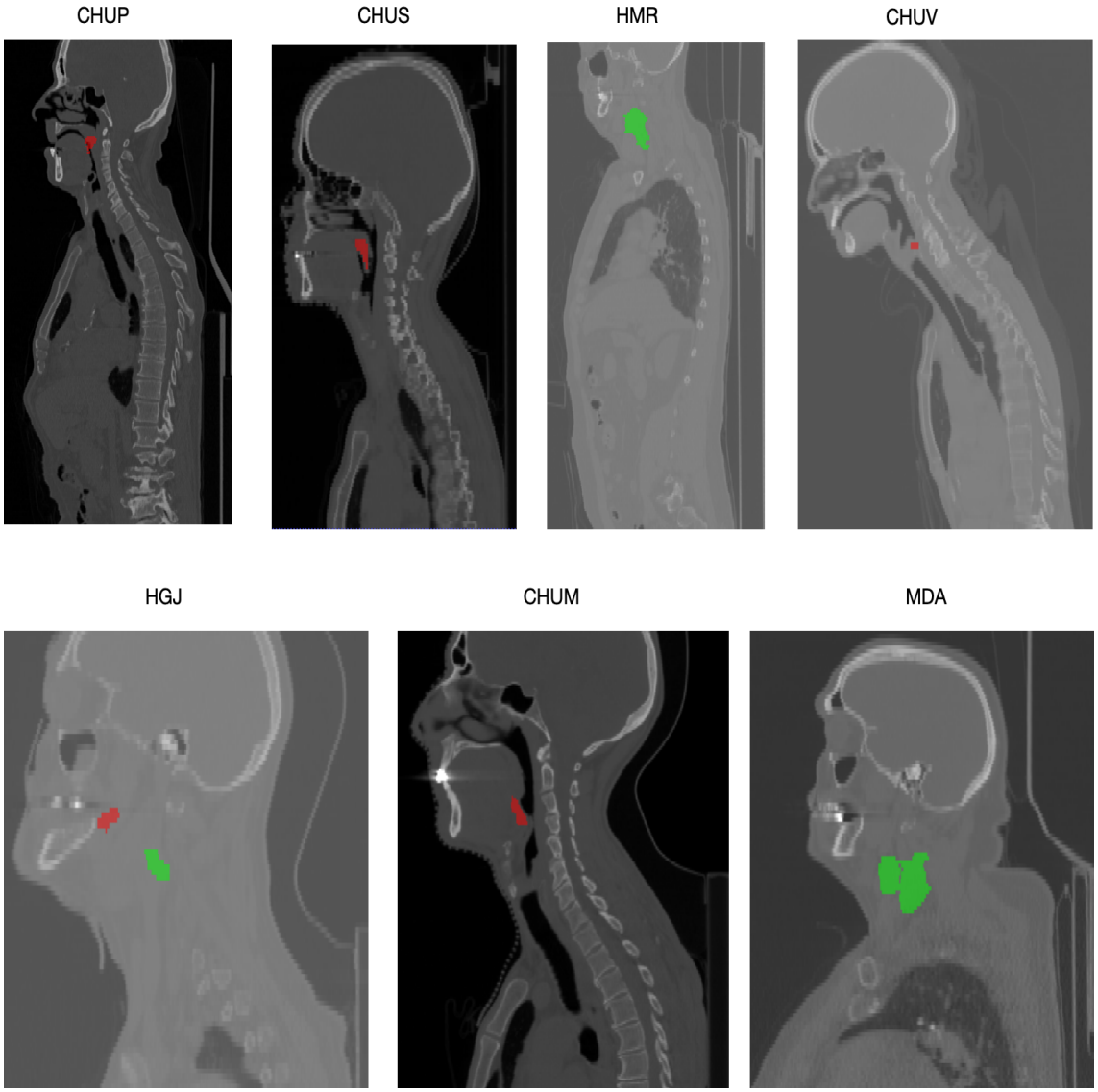

Figure 5.4: An example of a sagittal plane taken from each of the eight medical centers in the training dataset of the HECKTOR 2022 dataset highlighting the high variability in the image quality of the dataset. The GTVp is marked in red, and the GTVn is marked in green

Center	Acronym	Scanners	Training Cases	Testing Cases
Hôpital général juif, Montréal, Canada	HGJ	Discovery ST GE Healthcare	55	None
Centre hospitalier universitaire de Sherbooke, Sherbrooke, Canada	CHUS	GeminiGXL 16 Philips	72	None
Hôpital Maisonneuve-Rosemont, Montréal, Canada	HMR	Discovery STE GE Healthcare	18	None
Centre hospitalier de l'Université de Montréal, Montréal, Canada	CHUM	Discovery STE, GE Healthcare	56	None
Centre Hospitalier Universitaire Vaudois, Switzerland	CHUV	Discovery D690 TOF GE Healthcare	53	None
UniversitätsSpital Zürich, Switzerland	USZ	Discovery HR, RX, LS, TE, 690	None	101
Centre Henri Becquerel, Rouen, France	CHB	GE710 GE Healthcare	None	38
Centre Hospitalier Universitaire de Poitiers, France	CHUP	Biograph mCT 40 ToF GE Healthcare	72	None
MD Anderson Cancer Center, Houston, Texas, USA	MDA	Discovery HR, RX, ST, STE	198	200
Total			**524**	**359**

Table 5.2: The HECKTOR 2022 dataset [7] consists of 883 cases (524 for training and 359 for testing) collected from 9 medical centers using 12 different scanners across 4 different countries. The test dataset was collected from 3 different medical centers, of which 2 were not used in the training set.

5.4.2 Training and Testing

MMIS-Net

In the MMIS-Net all datasets were combined into a one-hot label space as demonstrated in Table 5.3. This approach effectively handles annotations present in one dataset but missing in another. For instance, in this work, there are two different pancreas datasets: [129], which includes segmentations for the pancreas and pancreas tumor, and [142], which includes segmentations only for the pancreas. The one-hot label space efficiently separates these as different labels without overlap. During training, MMIS-Net leverages the synergistic potential of one dataset to improve the performance of the other and vice versa. It also supports overlapping target structures, such as vessels or cancer classes within an organ, and retains different annotation protocol characteristics for the same target structure. During training, the following parameters were used: the learning rate was set to 0.1, the optimizer was Adam [46], the maximum epoch was set to 1000, the sigma parameters were fixed, and early stopping was used to avoid overfitting. The loss function used was a combination of cross-entropy and Dice loss. Here we aimed to improve the segmentation and detection performance on retinal OCT fluids. For this, we trained the algorithm by combining the 1337 publicly available volumes of the training sets of all 10 datasets and evaluated the results on the 42 volumes of the hidden test set of the RETOUCH [21] dataset set. Three evaluation metrics were used: Dice Score (DS): This measures the overlap between the predicted and ground truth segments, calculated as twice the intersection divided by the union. It ranges from 0 to 1, with 1 being the perfect score and 0 being the worst. Absolute Volume Difference (AVD): This is the absolute difference between the predicted and ground truth volumes. The value ranges from 0 to 1, with 0 being the best result and 1 being the worst. Area Under the Curve (AUC): This measures the ability of a binary classifier to distinguish between classes. The AUC score ranges from 0 to 1, with 1 being the perfect score and 0 being the worst. The DS and AVD were used to evaluate the segmentation of the retinal fluids on OCT scans, while the AUC was used to evaluate the detection of fluids on the retinal OCT scans. For fair comparison, we used the DS, AVD, and AUC evaluation metrics as they were the same evaluation metrics used by the organizers of the RETOUCH grand challenge for the retinal OCT dataset. Submissions are sent to the organizers, and the results are published on the challenge website and also sent to the teams. Submissions are limited to a maximum of three per team. The experimental setup was the same for all the experiments. The algorithm was written in Python using PyTorch backend libraries.

CVD_Net

The CVD_Net was trained for maximum of 1000 epochs with early stopping [107] to avoid over-fitting. Adam was the optimizer, and the learning rate was set to 0.01.The sum of the cross-entropy and Dice loss was taken as the loss function. For a fair comparison of the performance of the CVD_Net with other SOTA algorithms, our model was evaluated on a blind test set on the organizer's website. The ground truth for this test set is held by the organizers and is not available to the public. The hidden test set includes data from three medical centers, two of which are new centers not used during training. CVD_Net was trained on the entire training

Assigned Value	Region
0	Background
1	Liver
2	Liver tumor
3	Pancreas
4	Pancreas tumor
5	Hepatic vessels
6	Hepatic vessels tumor
7	Lung tumor
8	Spleen
9	Colon cancer
10	Bladder
11	Ulterus
12	Rectum
13	small bowel
14	Pancreas
15	Kidney
16	Kidney tumor
17	Intraretinal Fluid (IRF)
18	Subretinal Fluid (SRF)
19	Pigment Epithelium Detachments (PED)

Table 5.3: Table summarizing the labeling of the datasets in the one-hot label space. The segmentation tasks are labeled from 0 to 19.

set of 524 volumes and tested on 359 volumes of the hidden test set from 3 medical centers of which two were not seen at training. To further evaluate the generalization performance of the CVD_Net, the training dataset was split into two subsets: a training subset and a testing subset. The testing subset consists of data from medical centers not used in the training subset. The evaluation metric used was the Dice Score (DS), which is twice the intersection divided by the union. It measures the overlapping of the pixels, ranging from 0 to 1, with 1 being the perfect score and 0 being the worst. DS was the evaluating metric used by the challenge organizers, so for fair comparison, we have used DS.

The Friedman test was used to detect differences in performance between the algorithms evaluated on the segment classes per dataset. We computed the Friedman test statistic to check for significant differences and ranked the algorithms per segment class.)

The models were implemented in Python using the PyTorch library, and were trained on a GPU server with NVIDIA RTX A6000 48GB.

5.4.3 Results

This section presents a comparison of our results with other state-of-the-art (SOTA) architectures. Both methods were evaluated on the hidden test sets of the RETOUCH and HECKTOR 2022 challenges. For the MMIS-Net, we employed the Friedman test to assess the statistical significance of the algorithms' performance based on the combination of all three metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC). The results, including algorithm rankings and scores, are presented in Table 5.6. Also, the results for both algorithms are published online on the respective challenge websites and are compared with the SOTA models or leading teams in each competition. RETOUCH wesite[1] HECKTOR website[2] .

Multi-organ

From the experimental results we observed the following:

1. The MMIS-Net outperformed the SOTA algorithms on the segmentation task with a clear improvement in both DS and AVD, obtaining a mean of 0.83 and 0.035, respectively, on the RETOUCH retinal OCT hidden test set.

2. The MMIS-Net obtained the best DS score in all three fluid classes and the best AVD in two out of the three classes for the segmentation task on the RETOUCH retinal OCT hidden test set.

3. The MMIS-Net achieved a perfect AUC score of 1 alongside two other SOTA algorithms for the detection task on the RETOUCH retinal OCT hidden test set.

4. CVD_Net obtained the best mean AVD of 0.031 on the RETOUCH retinal OCT hidden test set.

5. CVD_Net obtained the best AVD of 0.032 for the segmentation of the PED fluid on the RETOUCH retinal OCT hidden test set.

6. For the RETOUCH retinal OCT segmentation and detection tasks, as well as the segmentation task, we notice a constant and steady high performance of the MMIS-Net algorithm, highlighting its robustness and generalizability.

7. The Friedman test on the combination of all 3 metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC) revealed statistically significant differences between at least two algorithms, with a p-values of 0.0041 and a Friedman test statistic of 27.3192.

Segmentation measured in DS and AVD on the RETOUCH retinal OCT hidden test set is highlighted in Table 5.4, and the detection task measured in AUC is highlighted in Table 5.5, with their corresponding bar charts in Figure 5.5 and Figure 5.6, respectively. To further demonstrate the high performance of the MMIS-Net, a visualization comparison of the predicted output of 5-fold cross validation on the RETOUCH training dataset is demonstrated in Figure 5.7.

[1] `https://retouch.grand-challenge.org/Home/`
[2] `https://hecktor.grand-challenge.org/Overview/`

Methods/Teams	Dice Score (DS)				Absolute Volume Difference (AVD)			
	IRF	SRF	PED	Avg.	IRF	SRF	PED	Avg.
MMIS-Net	**0.85**	**0.81**	**0.83**	**0.83**	**0.018**	**0.015**	0.071	0.035
nnUNet_RASPP	0.84	0.80	**0.83**	0.823	0.023	0.016	0.083	0.041
nnU-Net	**0.85**	0.78	0.82	0.817	0.019	0.017	0.074	0.036
CVD_Net	0.78	0.77	0.83	0.79	0.039	0.021	**0.032**	**0.031**
SFU	0.81	0.75	0.74	0.78	0.030	0.038	0.139	0.069
SAMedOCT	0.77	0.76	0.82	0.78	0.042	0.020	0.033	0.032
IAUNet_SPP_CL	0.79	0.74	0.77	0.77	0.021	0.026	0.061	0.036
UMN	0.69	0.70	0.77	0.72	0.091	0.029	0.114	0.078
MABIC	0.77	0.66	0.71	0.71	0.027	0.059	0.163	0.083
SVDNA	0.80	0.61	0.72	0.71	–	–	–	–
RMIT	0.72	0.70	0.69	0.70	0.040	0.072	0.1820	0.098
RetinAI	0.73	0.67	0.71	0.70	0.077	0.0419	0.2374	0.118
Helios	0.62	0.67	0.66	0.65	0.0517	0.055	0.288	0.132
NJUST	0.56	0.53	0.64	0.58	0.1130	0.0968	0.248	0.153
UCF	0.49	0.54	0.63	0.55	0.2723	0.1076	0.2762	0.219

Table 5.4: Performance evaluations of methods/teams, grouped by segmented classes and averages (Avg.), on the hidden test set of the RETOUCH grand challenge, measured in Dice Score (DS) and Absolute Volume Difference (AVD).

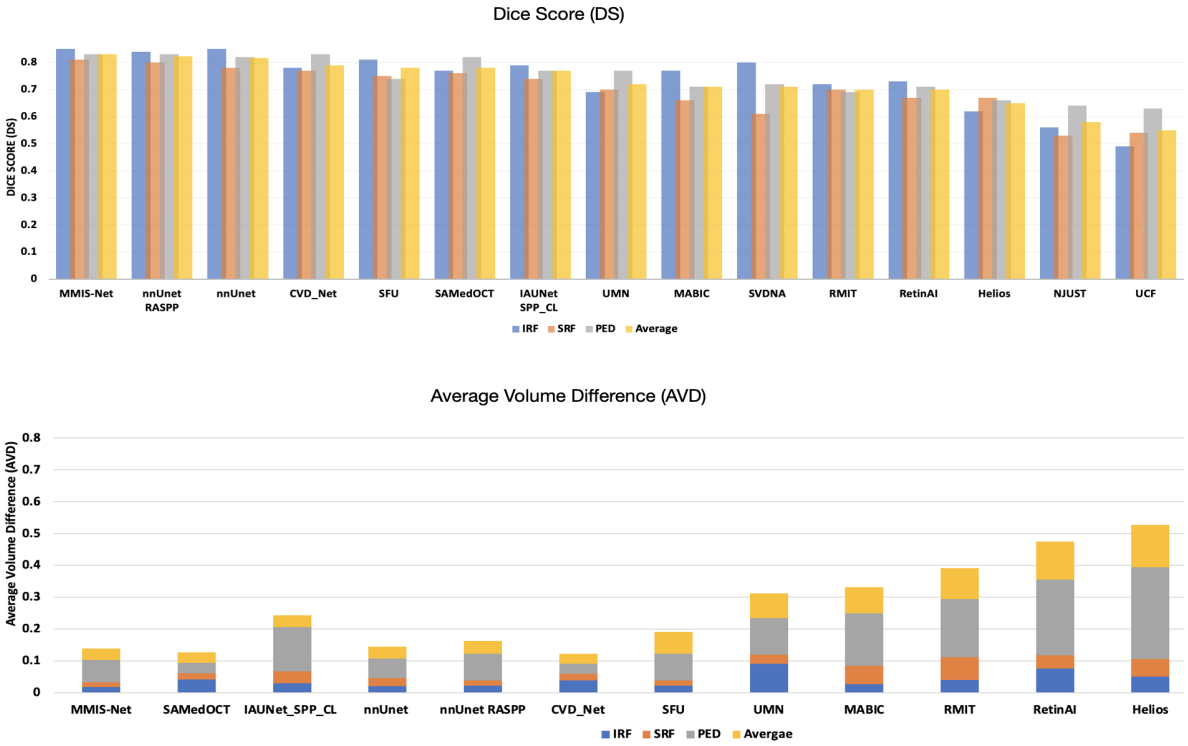

Figure 5.5: Comparison of performance evaluations for methods/teams, categorized by segmented classes and averages (Avg.), on the hidden test set of the RETOUCH grand challenge, measured with Dice Score (DS) and Absolute Volume Difference (AVD), presented in bar charts.

Methods	IRF	SRF	PED	Avg.
MMIS-Net	**1.0**	**1.0**	**1.0**	**1.0**
nnUNet	**1.0**	**1.0**	**1.0**	**1.0**
SFU	**1.0**	**1.0**	**1.0**	**1.0**
nnUNet_RASPP	0.93	0.97	**1.0**	0.97
CVD_Net	0.92	0.96	**1.0**	0.96
Helios	0.93	**1.0**	0.97	0.97
UCF	0.94	0.92	**1.0**	0.95
MABIC	0.86	**1.0**	0.97	0.94
UMN	0.91	0.92	0.95	0.93
RMIT	0.71	0.92	**1.0**	0.88
RetinAI	0.99	0.78	0.82	0.86
NJUST	0.70	0.83	0.98	0.84

Table 5.5: Evaluation performance of the fluids detection, measured in Area Under the Curve (AUC), grouped by segmented classes with their averages in columns and teams in rows on the hidden test set of the RETOUCH grand challenge.

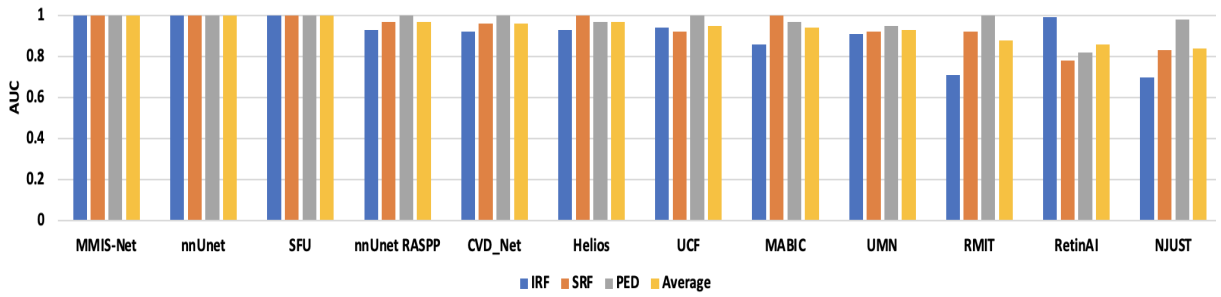

Figure 5.6: Performance evaluation of fluid detection, measured by Area Under the Curve (AUC), categorized by segmented classes and their averages, and grouped by teams on the hidden test set of the RETOUCH grand challenge.

Number of segment classes: 3
Number of algorithms : 12
Degrees of freedom: (11, 2)
Significance level (alpha): 0.05
p-value: 0.0041
Friedman statistic: 27.3192
Hypothesis: Alternative Hypothesis
Significant: There is a significant difference between at least two algorithms (p-value < 0.05).

Rank	Algorithm	Ranking Score
1	MMIS-Net	1.33
2	nnU-Net	2.33
3	nnUNet_RASPP	2.67
3	CVD_Net	2.67
4	SFU	3.67
5	UMN	6.00
6	MABIC	6.33
7	Helios	7.00
8	RMIT	7.67
9	RetinAI	8.33
10	UCF	9.00
11	NJUST	10.00

Table 5.6: The ranking (from best to worst) of the teams/algorithms based on the combination of all 3 metrics: Dice Score (DS), Average Volume Difference (AVD), and Area Under the Curve (AUC), using the Friedman test (a non-parametric test) indicates a significant difference between at least two of the algorithms, with a p-value of $0.0041 < 0.05$ and a Friedman test statistic of 27.3192.

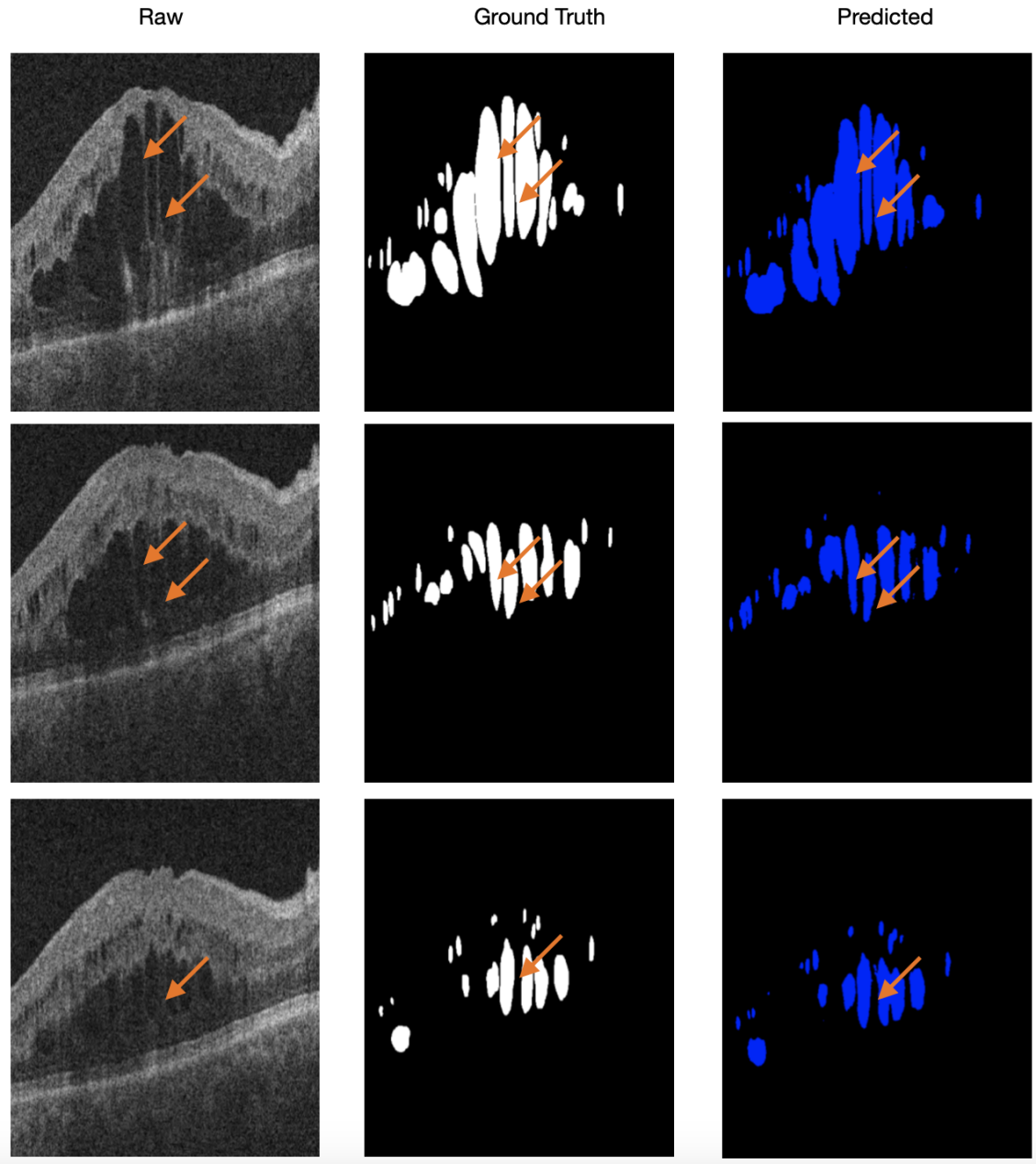

Figure 5.7: A visualization of B-Scans demonstrating the performance of MMIS-Net on the training set of the Retouch dataset using a 5-fold cross-validation. Orange arrows highlight details captured by MMIS-Net.

HECKTOR 2022

Evaluation results on the hidden HECKTOR 2022 testing dataset from from three medical centers, of which two medical centers are new and not in the training set, show that CVD_Net obtained a mean dice score of 0.77492 (0.77603 for GTVp and 0.77382 for GTVn) and MMIS-Net achieved a mean Dice score of 0.7734 (0.7740 for GTVp and 0.7737 for GTVn) as demonstrated in Table 5.7 with the corresponding bar chart in Figure 5.8. To further illustrate the generalization ability of our algorithm, we trained the CVD_Net on a subset of the training dataset and evaluated the performance on a holding subset from independent medical center not seen during training. Additionally, we provide comparisons of our proposed CVD_Net to other SOTA specialized and foundational models in this domain. This is illustrated in Table 5.8. B-Scans of the coronal view showing the raw data, annotated/ground truth, and corresponding predictions from different models are illustrated in Figure 5.9, demonstrating the slight performance advantage of CVD_Net.

Methods/Teams	GTVp	GTVn	Mean
NVIDIA(Nvauto) [133]	**0.80066**	**0.77539**	**0.78802**
CVD_Net	0.77603	0.77382	0.77492
MMIS-Net	0.77340	0.77400	0.77370
nn-UNet [84]	0.77485	0.76938	0.77212
MA-SAM [31]	0.67052	0.74453	0.70753

Table 5.7: Segmentation table of the Dice Scores (DS) by segment classes: primary tumors (GTVp) and Gross Tumor Volumes (GTVn) in columns, and algorithms/teams in rows. The evaluation performance by training on the entire training set from six medical centers and testing on the holding testing set from three medical centers, including two new independent medical centers not included in the training set.

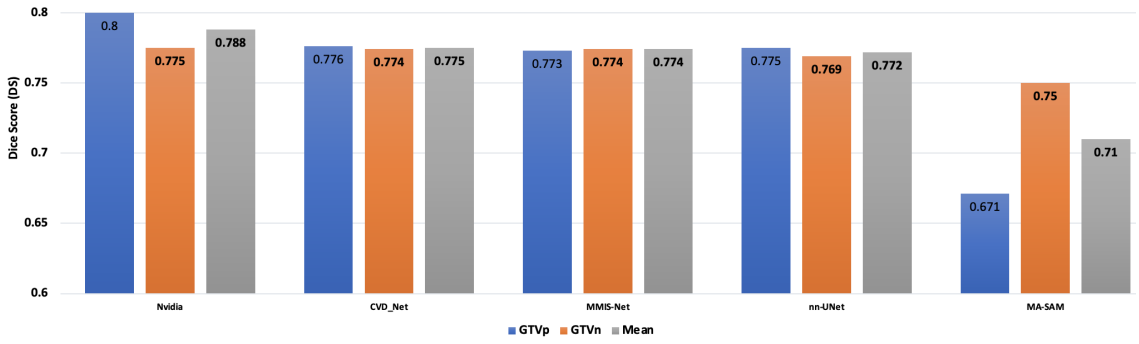

Figure 5.8: A visualisation comparison measured in Dice Scores (DS) by segment classes: primary tumors (GTVp) and Gross Tumor Volumes (GTVn), grouped by algorithms/teams. The evaluation performance by training on the entire training set from six medical centers and testing on the holding testing set from three medical centers, including two new independent medical centers not included in the training set.

Methods	**Training**	**Testing**	**GTVp**	**GTVn**	**Mean**
CVD_Net	CHUM, CHUP, CHUS, CHUV, MDA, HGJ	HMR	**0.7628**	**0.7781**	**0.7705**
nnUNet [84]	CHUM, CHUP, CHUS, CHUV, MDA, HGJ	HMR	0.7598	0.7758	0.7678
MA-SAM [199]	CHUM, CHUP, CHUS, CHUV, MDA, HGJ	HMR	0.5718	0.5879	0.5799
CVD_Net	CHUM, CHUP, CHUS, CHUV, MDA, HMR	HGJ	**0.7891**	**0.7634**	**0.7763**
nnUNet [84]	CHUM, CHUP, CHUS, CHUV, MDA, HMR	HGJ	0.7807	0.7597	0.7702
MA-SAM [199]	CHUM, CHUP, CHUS, CHUV, MDA, HMR	HGJ	0.6710	0.5781	0.6255
CVD_Net	CHUM, CHUP, CHUS, CHUV, MDA, HGJ	CHUV	**0.7781**	**0.7672**	**0.7727**
nnUNet [84]	CHUM, CHUP, CHUS, CHUV, MDA, HGJ	CHUV	0.7719	0.7596	0.7658
MA-SAM [199]	CHUM, CHUP, CHUS, HGJ, MDA, HMR	CHUV	0.6212	0.5949	0.6081

Table 5.8: A table comparing the generalizability performance of segmentation in Dice Scores (DS) by segment classes (columns) and algorithms (rows) for training on the training subset from five medical centres and testing on the holding testing set from an independent centre not seen during training.

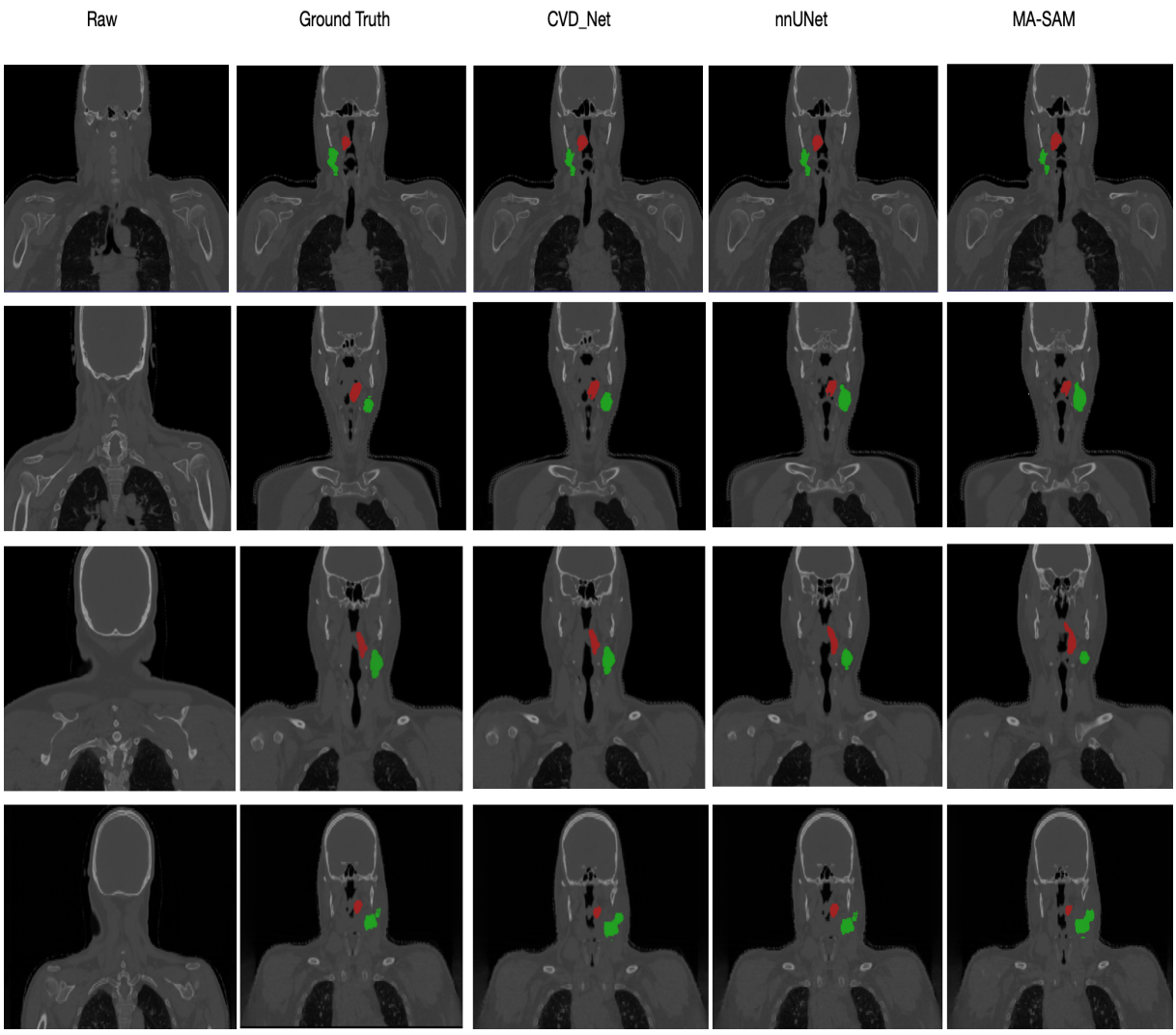

Figure 5.9: Coronal planes visualization comparing predictions from different architectures to the ground truth/human annotations and raw images. The GTVp is marked in red, and the GTVn is marked in green.

5.5 Summary

In this chapter, we have investigated the problem of knowledge transfer by combining datasets from multiple data sources, modalities, organs, and disease types. We used domain knowledge and similarity knowledge adapters to combat the problems of negative knowledge transfer and generalizability. We propose two novel algorithms which are MMIS-Net, and CVD_Net.

1) MMIS-Net is designed to segment multiple lesions from various organs across diverse image modalities using a single model. To address the issue of negative knowledge transfer, MMIS-Net introduces Similarity Fusion Blocks within its architecture. These blocks utilize supervised and selective knowledge transfer for feature map fusion at the pixel level, effectively reducing irrelevant and noisy signals in the output. Additionally, we efficiently created a one-hot label space to address the inconsistent class definitions and label contradictions from diverse modalities and body regions.

2) CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-Specific Batch Normalization), which combines CNNs for feature extraction, Vision Transformers for capturing long-range dependencies, and domain-specific adapters, to extract domain specific features and share common features across domain, reducing negative knowledge transfer thereby improving the overall model's generalization ability.

Both algorithms were evaluated on the multi-organ and HECKTOR 2022 datasets. Results on the hidden test sets of the RETOUCH and HECKTOR challenges show that:

1. MMIS-Net achieved a top mean Dice score (DS) of 0.83 and an absolute volume difference (AVD) of 0.035 for the retinal fluids segmentation task, along with a perfect Area Under the Curve (AUC) of 1 for the fluid detection task, outperforming state-of-the-art, specially designed algorithms and large foundation models for medical image segmentation on the RETOUCH hidden test set.

2. On the HECKTOR hidden test set, MMIS-Net achieved a mean Dice score of 0.774, which is comparable to state-of-the-art, specially designed models and outperforms large foundation models for medical image segmentation by a clear margin.

3. CVD_Net achieved the best mean absolute volume difference (AVD) of 0.031 on the RETOUCH retinal OCT hidden test set..

4. CVD_Net achieved the best absolute volume difference (AVD) of 0.032 for the segmentation of PED fluid on the RETOUCH retinal OCT hidden test set.

5. CVD_Net achieved a mean Dice score of 0.77492 on the RETOUCH hidden test set, which is comparable to specifically designed state-of-the-art architectures and exceeds the performance of large foundation models for medical image segmentation by a clear margin.

6. We have demonstrated the high generalizability of the CVD_Net by training on a subset from the training set, comprising data from six centers, and testing it on data from a new center (a holding subset of the training dataset).

We achieved state-of-the-art (SOTA) performance, surpassing that of large foundation models while using fewer resources.

7. We have demonstrated that while large foundation models, show promising generalization performances for this specific problem, specifically tailored deep networks such as MMIS-Net, CVD_Net, and nnUNet still offer a slight advantage for addressing these particular problems.

We believe the superior performance of both algorithms can be attributed to the following factors:

MMIS-Net: The integration of two key features into the CNN backbone: (i) Similarity Fusion blocks for supervision and similarity-based knowledge selection, which enhance feature map fusion, and (ii) a one-hot label space to address inconsistent class definitions and label contradictions. This label space allows for handling of classes that are present in one dataset but absent in another while preserving distinct annotation protocol characteristics for the same target structure during training.

CVD_Net: The use of a hybrid backbone combining a Convolutional Neural Network (CNN) for feature extraction and a Vision Transformer (ViT) to capture long-range dependencies, along with Domain-Specific Batch Normalization to address negative knowledge transfer.

MMIS-Net and CVD_Net complement each other and should be used in different scenarios for future research depending on the size of the dataset. MMIS-Net is built on a CNN backbone, and CNNs are known for their ability to capture local contextual features [51]. Therefore, it would be suitable for projects with small or medium-sized datasets. On the other hand, CVD_Net is a hybrid combination of CNN and ViT (Transformers). Transformers are known for their ability to capture long-range dependencies [51] and would be suitable for projects with very large datasets.

Chapter 6

Discussion and Conclusion

6.1 Introduction

This work has presented novel approaches to enhance the detection, segmentation, and generalization of diseases in medical images, focusing on both specifically designed architectures and universal/general architectures. This chapter evaluates the methods, contributions, and outcomes of these approaches, with reference to the research objectives outlined in Section 1.1.

This chapter is organized as follows: Section 6.3 provides an overview of the research, summarizing the introduction from Chapter 1, the literature review from Chapter 2, and the key contributions and innovations presented in Chapters 3, 4, and 5. Next, Section 6.5 highlights the practical implications of the work. Finally, Section 6.4 outlines the limitations of the study, followed by Section 6.6, which offers suggestions for future research directions.

6.2 Summary of the Thesis and Main Findings

Deep learning methods have been successful in the segmentation and detection of diseases in medical images. However, most of these methods are trained and tested on data from the same sources, hence fail to generalize to new, unseen data like in real-world scenarios. One of the main reasons for this limitation is the domain shift between training and testing datasets. A potential solution to this problem is to develop a single, universal, and generalizable model by combining data from diverse sources, modalities, organs, and disease types. However, simply merging these diverse datasets can lead to another challenge known as Negative Knowledge Transfer, where knowledge gained from one domain or dataset negatively impacts others, ultimately degrading the overall model performance. In this thesis, we propose novel architectures to enhance the generalization performance of deep learning models for the segmentation and detection in medical image and it is summarized as follows:

First Chapter 1 introduces this work, outlining its aim and objectives, the thesis structure, the significance of the study, its practical implications, scope, and data collection process.

Chapter 2 provides a comprehensive review of recent advancements in the field, categorized into 6 main areas: Specific Models, Domain Adaptation, Universal Model, Federated Learning, Fine-tuning, and Foundation models. Additionally, it

summarizes several large publicly available annotated medical image datasets from various sources.

Our first contribution is presented in Chapter 3 : Enhancing Retinal Disease Detection, Segmentation, and Generalization with an ASPP Block and Residual Connections Across Diverse Data Sources. Here, we proposed a novel architecture termed nnUNet_RASPP, for the detection, segmentation, and generalization of 3 retinal diseases from diverse data sources collected using 3 different manufacturer devices. nnUNet_RASPP was evaluated on the RETOUCH challenge dataset [21], and experimental results on the hidden test set demonstrate that nnUNet_RASPP outperforms current state-of-the-art (SOTA) architectures, including large foundation model for medical image segmentation by a clear margin. nnUNet_RASPP is currently the winner of both the online and offline versions of the challenge. The work is published in [136].

In Chapter 4 we present our second contribution : Dynamic Network for Global Context-Aware Disease Segmentation in Retinal Images Using Multiple ASPP and SE Blocks. In this chapter, we propose a novel architecture called Deep_ResUNet++. The algorithm was evaluated on 2 benchmark datasets: the Annotated Retinal OCT Images (AROI) [131] and the Duke DME [40] dataset. Experimental results show that Deep_ResUNet++ outperforms the current SOTA algorithms by a clear margin on these benchmarks. This work presented is published in [140], and [139].

Our third and final contribution is presented in Chapter 5:Enhancing Medical Image Segmentation Through Knowledge Transfer with Domain-Specific Adapters Across Diverse Data Sources. Here, we propose two novel algorithms: a pure Convolutional Neural Network (CNN) called MMIS-Net and a hybrid model combining CNN and Vision Transformer (ViT) called CVD_Net. Both architectures were evaluated on two sets of datasets. The first set consists of 19 benchmark datasets across two modalities, while the second set is the HEad and neCK TumOR (HECKTOR) challenge 2022 dataset [7] collected from 9 medical centers across the world for 2 modalities. Experimental results on the hidden test set of the RETOUCH Grand Challenge dataset show that MMIS-Net outperforms the current SOTA algorithms, including large foundation models for medical image segmentation, by a clear margin. Also, results on the hidden test set of the HECKTOR dataset indicate that CVD_Net achieves performance comparable to SOTA algorithms. The CVD_Net is published in [137], while MMIS-Net is currently under review for journal publication.

6.3 Contributions

This research focuses on enhancing disease/pathogen detection, segmentation, and generalization in medical images. The key contributions are summarized as follows:
Enhancing Retinal Disease Detection, Segmentation, and Generalization with an ASPP Block and Residual Connections Across Diverse Data Sources:
Chapter 3 investigates the potential of capturing global contextual features at varying rates to enhance segmentation and generalization performance in a specifically designed architecture. This is achieved by incorporating an Atrous Spatial Pyramid Pooling (ASPP) block just before the input layer within the nn-UNet architecture [84], resulting in a novel algorithm termed nnUNet_RASPP. The ASPP was used to : (i) effectively capture structures of varying sizes within the images, (ii) adapt more effectively to different dataset characteristics, such as variations in resolution and noise, (iii) capture both local and global context, and (iv) reduce the model's over-reliance on features from any single scale. Adding residual connections to address overfitting. Additionally, we conducted a performance evaluation of the top teams in the RETOUCH challenge, highlighting the different architectures employed. The nnUNet_RASPP was evaluated on the MICCAI 2017 RETOUCH Grand Challenge benchmark dataset [21], which was acquired from multiple sources using three vendor devices. Experimental results on the hidden test set demonstrated that nnUNet_RASPP outperformed state-of-the-art specific designed architectures and large foundation models for medical image segmentation by a clear margin. Additionally, nnUNet_RASPP exhibited excellent generalization performance on new, unseen data. The results are published on the organizer's website [1], and we are the current winners of both the online and offline challenge.
Dynamic Network for Global Context-Aware Disease Segmentation in Retinal Images Using Multiple ASPP and SE Blocks :
Chapter 4 explores the potential of capturing global contextual features at varying rates to enhance the segmentation and generalization of diseases/pathogens in medical images, to propose a novel algorithm, called Deep_ResUNet++. The Deep_ResUNet++ integrates multiple Atrous Spatial Pyramid Pooling (ASPP) and Squeeze-and-Excitation (SE) blocks at various locations to effectively capture global contextual features. Additionally, residual connections are incorporated in both the encoding and decoding paths to address the vanishing gradient problem. The architecture is built on a dynamic convolutional neural network (CNN) backbone that adjusts its kernel size and network depth based on the input, providing enhanced adaptability and performance. The use of multiple ASPP and SE blocks in a CNN segmentation network offers a more detailed and effective method for feature extraction, context aggregation, and feature recalibration. Deep_ResUNet++ was evaluated on two benchmark datasets: the Annotated Retinal OCT Images (AROI) dataset and the Duke DME dataset, which were collected from patients with two different disease types. Experimental results demonstrate that Deep_ResUNet++ significantly outperformed current state-of-the-art architectures by a clear margin. On the AROI dataset, Deep_ResUNet++ achieved a mean Dice score of 0.98, surpassing the second-best model by 0.01 (1%) and consistently scoring above 0.90 across all classes. On the Duke DME dataset, it achieved a mean Dice score of 0.88,

[1] `https://retouch.grand-challenge.org/Home/`

outperforming the second-best model by 0.02 (2%).

Enhancing Medical Image Segmentation Through Knowledge Transfer with Domain-Specific Adapters Across Diverse Data Sources:
Chapter 5 addresses the challenge of limited annotated medical datasets by combining multiple small annotated datasets from various sources, modalities, and disease types to build a unified model with high generalizability. We propose two novel algorithms utilizing knowledge transfer and domain-specific adapters:
1) MMIS-Net (MultiModal Medical Image Segmentation Network): This approach tackles label inconsistency from multiple data sources by creating effective one-hot labels and incorporating similarity fusion blocks into the U-Net architecture.
2) CVD_Net (Convolutional Neural Network and Vision Transformer with Domain-Specific Batch Normalization): Building on the previous methods, CVD_Net integrates Domain-Specific Batch Normalization (DSBN) with a combination of CNN and Transformer architectures. The DSBN was used to capture and share domain-specific context within a shared parameter backbone, thereby reducing the transfer of negative knowledge and improving the model's generalization ability.

Both MMIS-Net and CVD_Net were trained on two groups of datasets. The first group consisted of 10 benchmark datasets covering 19 organs across two modalities, while the second group included the HECKTOR 2022 dataset, collected from 9 medical centers around the world. For a fair comparison, MMIS-Net and CVD_Net were evaluated on hidden test datasets. Experimental results demonstrated that both algorithms outperformed state-of-the-art, task-specific algorithms and large foundation models by a significant margin. Additionally, MMIS-Net and CVD_Net exhibited high generalizability on new, unseen data. The results are published on the respective challenge websites: RETOUCH wesite[2] and HECKTOR website[3] .
To effectively detect, diagnose, and monitor diseases and pathogens in medical images, it is crucial to understand human organ anatomy, structure, the changes caused by disease, imaging techniques, their effects, and their consequences. The appendix provides a brief review of the anatomical structures, diseases, and imaging techniques related to the primary organs studied.

[2] `https://retouch.grand-challenge.org/Home/`
[3] `https://hecktor.grand-challenge.org/Overview/`

6.4 Limitations

Despite the success and advancements of deep learning methods for medical image segmentation, and the contributions of this research, certain limitations remain, some of which are briefly discussed below.

1. **Network Dimensional Adaptability**: In Chapter 3, we introduced nnUNet_RASPP, a 3D algorithm that outperformed state-of-the-art 3D methods on the RETOUCH dataset. However, as demonstrated in Chapter 4, its performance did not translate effectively when evaluated on 2D B-Scans. We attribute this limitation to the lack of the third dimension in 2D B-Scans, which otherwise provides valuable spatial correlation between scans in volumetric images. Consequently, when transitioning from 2D to 3D, nnUNet_RASPP creates an artificial third dimension, which may introduce additional noise into the dataset, ultimately leading to poorer performance.

2. **Manual Parameter Setting**: In Chapter 4, we introduced Deep_ResUNet++, in which key hyper-parameters, such as batch size and learning rate, were set manually (trying an error). This approach does not guarantee optimal performance.

3. **Data Hunger**: In Chapter 5, we introduced CVD_Net, a hybrid approach that combined convolutional neural networks (CNNs) and vision transformers (ViTs). As demonstrated in [51], ViTs are more data-hungry than CNNs, and therefore, a sufficiently large dataset is required to combine both backbones efficiently.

4. **Limited Computational Resources**: Also, in Chapter 5, we introduced MMIS-Net, that combined data from diverse sources, modalities, organs, and disease types to significantly minimize the domain shift gap and increase the size of the training set. This, in turn, demands substantial computational power, which stretched our computational resources and limited our ability to train models on larger datasets.

5. **Imbalanced Dataset Size**: Furthermore, in Chapter 5, there was variation in the size of datasets within the multi-organ dataset, for training, smaller datasets were augmented to match the size of the larger ones. This process can introduce additional noise into the training dataset.

6. **Evaluation Metrics**: In Chapter 2, various evaluation metrics such as Dice Score (DS), Intersection over Union (IoU), Absolute Volume Difference (AVD), and accuracy are used to assess model performance. However, these metrics each rely on different evaluation techniques, and different researchers may prioritize different metrics. This lack of standardization complicates direct comparison between models, making it difficult to achieve a fair assessment of state-of-the-art algorithms.

7. **Dynamic Network Adjustment:** In Chapter 4, we introduced Deep_ResUNet++, a dynamic network that adapts based on the input image

shape and dataset size. This approach can be further improved by adjusting the network based on the disease.

8. **Generalization Beyond Evaluated Datasets:** Although the work presented in this thesis was trained on 11 benchmark datasets covering 20 segmentation tasks across 2 modalities, the architectures were tested on 2 datasets across 3 modalities. Expanding the testing datasets to include more structures and modalities would further validate and demonstrate the generalization performance of the models.

9. **Limited Image Modality:** The work presented in this thesis was evaluated on three imaging modalities: CT, PET, and OCT. Expanding the number of modalities to include others such as X-ray, MRI, and ultrasound would further increase the model's diversity and enhance its generalization ability.

10. **Limited Test Submissions:** Two of the datasets used in this work were obtained from online competitions/challenges. These competitions provide a fair comparison platform for participants by keeping the labels of the testing datasets hidden from the public. However, one limitation of this approach is the restricted number of submissions allowed per participant, which limits the number of experiments that could be conducted for each dataset.

11. **Lack of Automation:** In this work, we have proposed 4 separate novel approaches. However, one limitation is the absence of a unified framework that integrates all 4 algorithms. Such a framework could automatically determine the most suitable algorithm for a given dataset or problem.

12. **Exploration of Other Backbones:** Although this work has reviewed and explored the two most predominantly used backbones in deep learning: CNN and ViT. It did not explore other backbones, such as Mamba[121] from state-space models, or hybrid combinations of Mamba, CNN, and ViT.

Addressing these limitations is crucial for further advancing in deep learning models in medical image segmentation and generalization.

6.5 Significance and Impacts

The models developed in this work can be directly applied to diagnose and monitor disease progression in medical images. Given their high performance and generalization capability, these models can assist in diagnosing diseases at early stages, even when they are not easily detectable by human experts. This makes them valuable decision-support tools, providing a reliable second opinion.

Also, the the Deep_ResUNet++ demonstrated an overall performance higher than the inter-observer variability (agreement between human expert annotators), indicating that the model can handle less complex, time-consuming tasks, allowing doctors to focus on more challenging cases.

In addition, the models require fewer computational resources compared to large foundation models, making them reliable, portable, and deployable. This ensures their applicability even in areas with limited internet access.

Furthermore, significant work has been done in the diagnosis and segmentation of diseases in medical imaging using deep learning methods. However, most of these approaches are task-specific, with limited emphasis on generalization across multiple and diverse data sources. This research addresses that gap by contributing models designed to generalize effectively across diverse datasets, organs, disease types, and modalities.

Our work is published on the public domain and provides a new benchmark for further research and comparison in the in the generalization and segmentation of diseases in medical images. In addition, we have compiled and shared resources and links to large multi-source datasets and public code bases for medical image segmentation tasks providing valuable resources for researchers in the field.

The models developed in this work have the potential to improve patient care by facilitating early and accurate disease detection while reducing the socio-economic burden on both patients and healthcare systems.

6.6 Future Research Directions

While the research presented in this thesis shows promising results and lays a strong foundation for further exploration, several key areas deserve focus in future work:

Self-Parameterize: In the future, we plan to transform the Deep_ResUNet++ into a self-parameterizing framework, addressing the issue of manual parameter setting (trying an error). This modification would enable the model to automatically select the optimal parameters for a given dataset.

Cropping The Region of Interest: One way to reduce space complexity is by reducing the size of the images. In the future, for the MMIS-Net network, we plan to incorporate a pre-processing step that identifies and crops the region of interest. This enhancement would ensure that only a subset of the image is used for training, thereby reducing the computational resources required for training.

Imbalanced Dataset Size: In MMIS-Net, to address the issue of imbalanced dataset sizes, we augmented the smaller datasets to match the size of the largest dataset using standard deep learning data augmentation techniques such as flipping, cropping, and rotation. However, these approaches can introduce additional noise into the dataset. In the future, we plan to leverage the presence of unlabeled datasets on public domain and use Generative Adversarial Networks (GANs) to generate labeled data from unlabeled datasets, thereby increasing the size of smaller datasets more effectively.

MMIS-Net with Domain Adaption: In Chapter 5, we introduced MMIS-Net, a CNN-based backbone architecture that combined multiple datasets into a single label space. In the future, we plan to explore the potential of handling each domain separately by incorporating domain-specific adapters into the network, with each per domain and a shared parameter space to share knowledge common across all domains. This approach would be similar to CVD_Net but without the inclusion of transformers.

Combining Image and Tabular Data: So far, all the models presented in this research have been trained exclusively on image data. However, most of the datasets used also include tabular data with attributes such as patient demographics, which could be valuable for training. In the future, we plan to explore the potential of integrating both tabular and image data for training to improve the model's performance.

New Test Cases : Broadening the scope of the testing dataset to include samples from different organs, modalities, and disease types would further validate the generalizability and robustness of our algorithms. In the future, we plan to evaluate our algorithms to more datasets once they are publicly available.

Appendix A

Overview of Anatomical Structures

To effectively diagnose and monitor diseases using medical images, it is essential to understand the general structure, anatomy, and morphological changes of the human body. This research focuses on two key anatomical regions: the eye, and head and neck. In this section, we provide a brief overview of the anatomy, morphology, and common pathological changes associated with these areas.

A.1 Human Eye Anatomy Overview

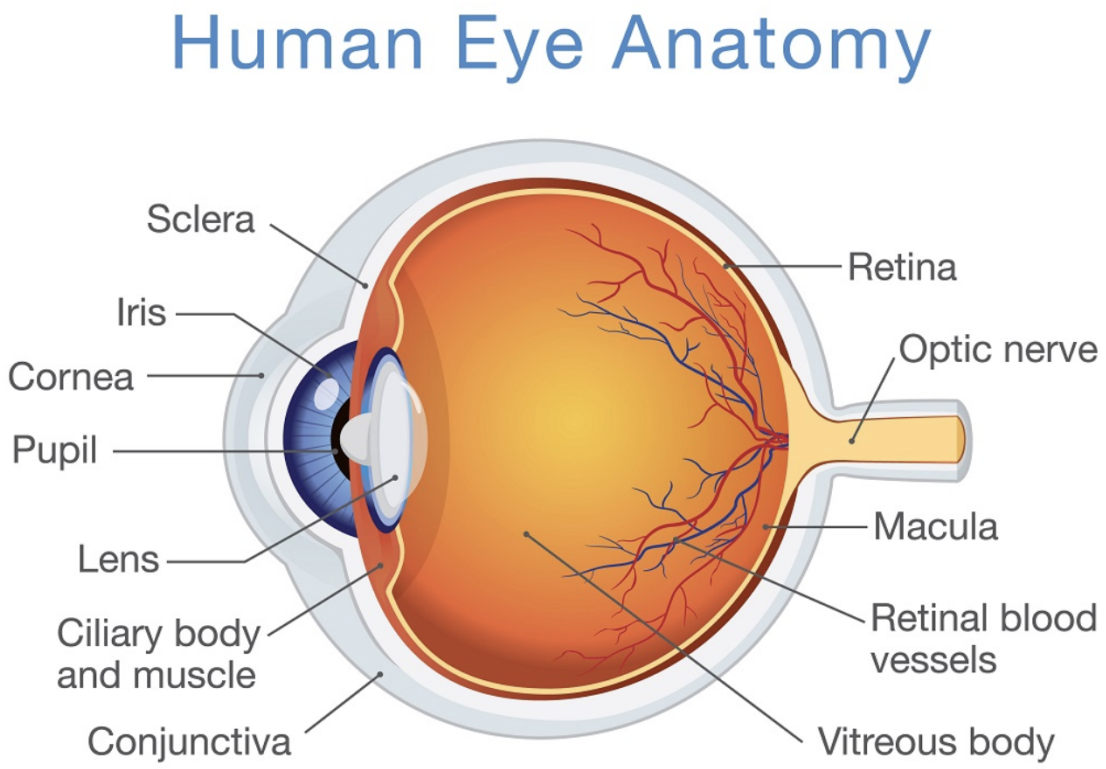

Figure A.1: An illustration depicting the primary components of the human eye. Image taken from [81]

Sight is one of the most important human senses, and the eye is the primary organ responsible for it. The human eye is composed of many parts, some of which are briefly discussed below:
Retina: The layer of cells at the back of the eye that converts light into electrical signals, which are sent to the brain, allowing us to see images.
Macula: The central area of the retina responsible for central vision.
Fovea: A small depression inside the macula where vision is the sharpest.
Choroid: Located between the retina and the sclera (the white part of the eye), it contains blood vessels that provide nutrients to the eye.
Optic Disc: The area where the retina connects to the optic nerve, which is critical for vision as it transmits visual information to the brain.
Blood Vessels: The blood vessels that support the inner retina. The central retinal artery and its branches supply the retina with oxygen and nutrients, while the central retinal vein and its branches remove carbon dioxide and waste.
Light rays entering the eye are focused on the retina, which senses the light and converts it to electrical impulses. These impulses are sent through the optic nerve to the brain, which interprets them as the images we see. The retina consists of hundreds of millions of neurons, many of which are photoreceptors that detect and respond to light. There are two types of photoreceptors: rods and cones. Rods detect motion and enable vision in dim lighting conditions. They are distributed throughout the retina, with a concentration along its periphery. Cones allow us to see color and detail; they are concentrated in the macula, a small area in the central part of the retina.
An illustration showing the main components and structure of the human eye is provided in Figure A.1.

Retinal Structure Analysis

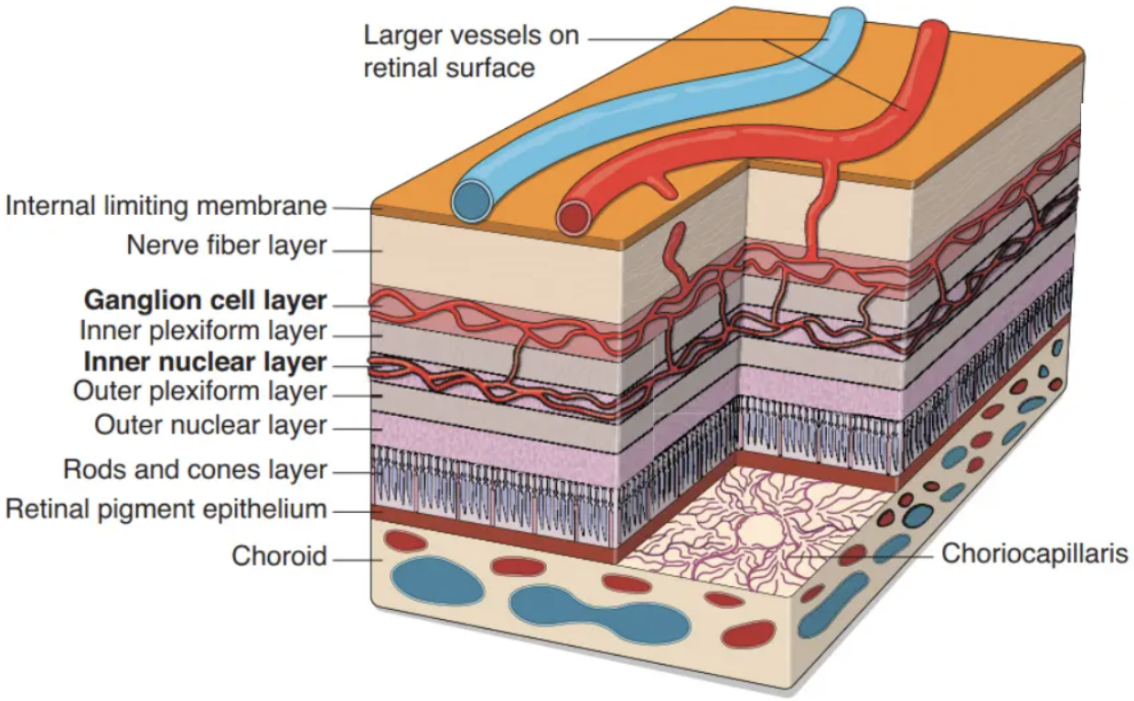

Figure A.2: A scan of the eye illustrating the retinal at the top and it's corresponding layers at the bottom.

The retina is a light-sensitive layer of nerve tissue at the back of the eye that receives images and sends them as electric signals through the optic nerve to the brain. The structure of the eye, as depicted in [155], shows the layers of the retina, as demonstrated in Figure A.2. The retina comprises several layers, generally classified into 10 distinct layers:
1) **Inner limiting membrane (ILM)**: The innermost surface bordering the neural retina and vitreous humor, containing astrocytes and the end feet of Muller cells.
2) **Nerve fiber layer (NFL)**: The second innermost layer of the retina, consisting of nerve fibers from the ganglion cells.
3) **Ganglion cell layer**: Contains the retinal ganglion cells (RGCs) and displaced amacrine cells.
4) **Inner plexiform layer**: Composed of a dense network of interlaced dendrites from RGCs and cells of the inner nuclear layer.
5) **Inner nuclear layer**: Contains the cell bodies of bipolar cells, horizontal cells, and amacrine cells.
6) **Outer plexiform layer (OPL)**: Features neuronal synapses between rods and cones and the footplates of horizontal cells.
7) **Outer nuclear layer (ONL)**: Contains the rod and cone granules that sense photons, as well as extensions from the rod and cone cell bodies.
8) **External limiting membrane (ELM)** : Includes the bases of the rod and cone photoreceptor cell bodies.
9) **Layer of rods and cones**: Contains the photoreceptor cells (rods and cones) themselves.
10) **Retinal pigmented epithelium (RPE)**: A single layer of cells tightly joined

to form a barrier between the retina and the underlying choroid, supporting the retina.

Retinal Diseases

Figure A.3: A scan of the retina illustrates fluid leakage affecting the retina due to neovascularization in age-related macular degeneration (AMD) at the top, vision loss in AMD at the bottom right, and vision loss in diabetic macular edema (DME) at the bottom left. Images taken from [154]

Many eye diseases manifest in the retina, including age-related macular degeneration (AMD) and diabetic macular edema (DME), both of which can severely impair vision. AMD and DME are leading causes of vision impairment in developed countries [108]. AMD predominantly affects older patients, while DME is common among working-age individuals.

Macular edema: This is the swelling of the central retina caused by leakage from the retinal capillaries and the subsequent accumulation of fluid within the intercellular spaces of the retina. This condition leads to sudden and severe vision loss and often occurs secondary to retinal diseases such as AMD and DME.

Age-related macular degeneration (AMD): This is an eye disease that blurs central vision. It primarily affects older adults and is a leading cause of vision loss in this demographic. While AMD does not cause complete blindness, it can severely impair central vision. AMD is painless and does not affect the appearance of the eyes. There are two types of AMD:

1) Dry AMD (atrophic AMD): This is an early stage of AMD which is more common, less severe, and progresses slowly over several years. It is characterized by the thinning of the macula and the presence of drusen (yellow deposits). A few small drusen might not affect your vision initially, but as they increase in size and number, they can cause vision to dim or become distorted.

2)Wet AMD (neovascular AMD): It is an advanced stage of AMD that is less common, more severe and leads to faster vision loss. It occurs when unstable blood vessels grow beneath the macula. These vessels leak blood and fluid into the retina, causing vision distortion. This condition can also result in blind spots and loss of central vision. Over time, the bleeding and abnormal blood vessels can form a scar, leading to permanent central vision loss.

Some of the symptoms of AMD include blurred or reduced central vision, difficulty recognizing faces, and a dark or empty area in the center of vision.

Diabetic macular edema (DME): This involves the accumulation of excess fluid in the extracellular space within the retina, causing swelling in the macular area. This condition leads to blurred vision and is common among diabetic patients and working-age adults.

The symptoms of DME can vary but commonly include: Blurred Vision: Especially in the central visual field. Distorted Vision: Straight lines may appear wavy. Color Perception Changes: Colors may appear washed out or less vibrant. Floaters: Small spots or strings floating in your vision. Difficulty Reading or Seeing Faces: Due to the central vision impairment.

Diabetic retinopathy(DR): This occurs when damaged blood vessels in the eye leak blood or fluid into the retina, a condition caused by the accumulation of sugar in the blood due to diabetes. DR severely impairs vision and is common among working-age adults. Initially, diabetic retinopathy may cause no symptoms or only mild vision problems, but it can eventually lead to blindness.

Diabetic retinopathy(DR):

1)Non-Proliferative Diabetic Retinopathy (NPDR): Early stage, where blood vessels weaken and leak fluid or blood.

2)Proliferative Diabetic Retinopathy (PDR): Advanced stage, where new, abnormal blood vessels grow, leading to severe vision problems.

Common symptoms of DR include: Spots or dark strings floating in vision, blurred vision, fluctuating vision, and vision loss.

Retinal detachment (RD): This occurs when the retina is pulled away from its normal position at the back of the eye. Although it can be repaired with surgery, prompt detection and treatment are crucial to prevent sight loss in the affected eye. There are three types of retinal detachment:
1)Rhegmatogenous: Caused by a tear or break in the retina.
2)Tractional: Occurs when scar tissue pulls on the retina.
3) Exudative: Fluid accumulates under the retina without a tear or break.
The most common symtoms of DR include, Sudden flashes of light, floaters, a shadow or curtain over a part of the visual field. An example demonstrating the changes in retinal structure caused by swelling and fluid leakage, along with vision loss in AMD and DME, is illustrated in Figure A.3.

Retinal Disease Treatment Options

Currently, there is no cure for these diseases. Treatment for retina diseases varies depending on the specific condition and its severity. Common treatments include:
Anti-VEGF Injections: Medications like ranibizumab, aflibercept, or bevacizumab are injected into the eye to inhibit vascular endothelial growth factor (VEGF), reducing abnormal blood vessel growth and leakage.
Corticosteroids: Injections or implants, such as dexamethasone or fluocinolone, can reduce inflammation and swelling.
Laser Photocoagulation: Uses laser energy to seal leaking blood vessels and reduce fluid accumulation.
Surgery: Vitrectomy might be necessary in severe cases to remove vitreous gel and scar tissue affecting the macula.
The effectiveness of these treatments depend on early diagnosis and effective monitoring of the disease progression. Early diagnosis allows doctors to advise patients on behavioral changes, such as change of diet and engaging in regular exercise, which can slow the disease's progression or even prevent it from advancing to more severe stages. Currently, much of this work is done manually, which is time-consuming and prone to error. Additionally, anti-VEGF drugs are expensive and must be administered frequently over an extended period, posing a socio-economic burden on both patients and the healthcare system. Hence the need to develop an automated tool for the diagnosis and monitoring of the disease progress.

Retinal Imaging Techniques

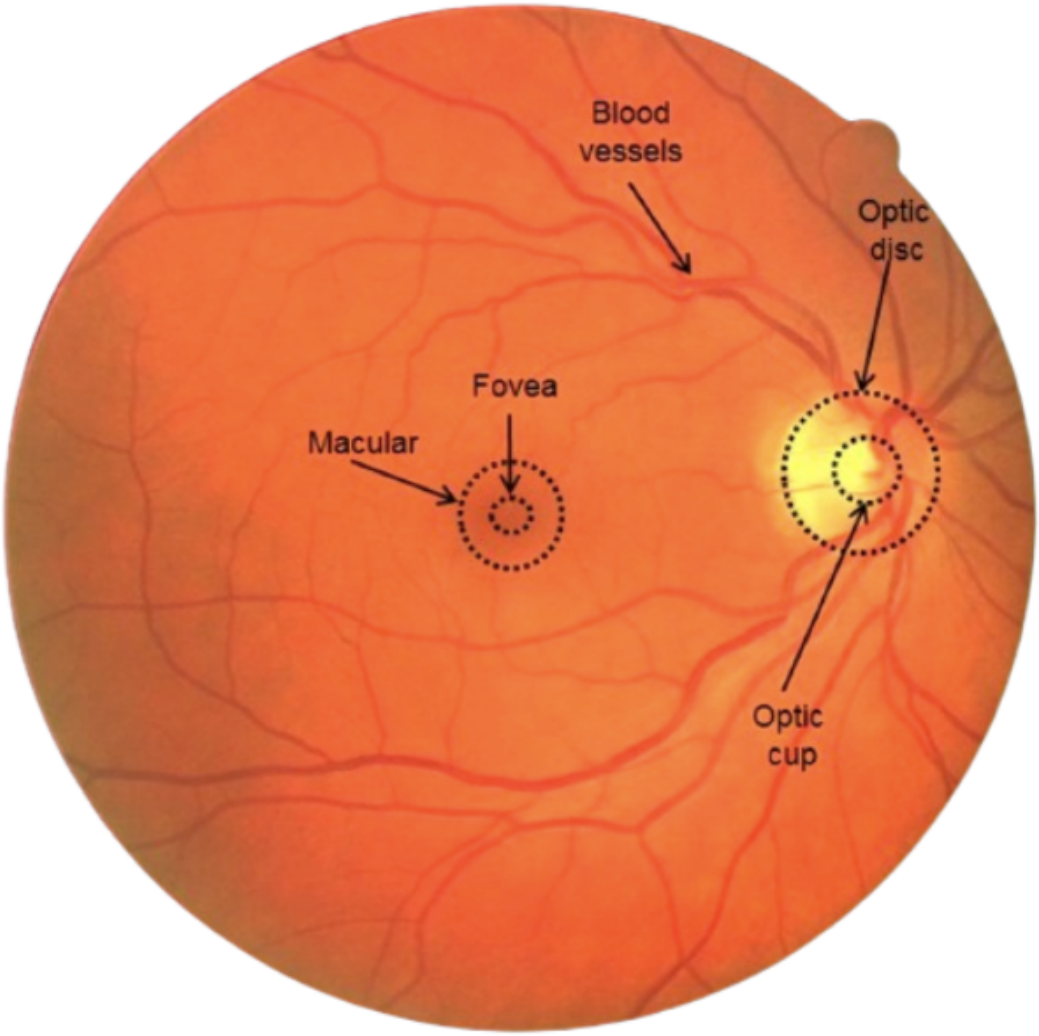

Figure A.4: A fundus photograph showing the Macular, Fovea, Blood vessels, optic disc and optic cup.

The two most commonly used retinal imaging techniques are fundus photography and optical coherence tomography (OCT).

Fundus photography: The back part of the inside of the eye is called the fundus. It is where the retina, macula, fovea, choroid and optic disc, as well as blood vessels, are located. Fundus photography employs multiple lenses and a camera to capture 2D images of the fundus. The first commercially available fundus camera produced by Carl Zeiss in 1926 [143].

Optical Coherence Tomography (OCT) : Optical Coherence Tomography (OCT) is a 3D, non-invasive imaging technique that provides detailed cross-sectional scans of the eye, revealing the retinal structure and anatomy and can be used to analyze and monitor the presence of pathogens or diseases in the retina. OCT uses light waves, typically in the near-infrared spectral range, to measure retinal thickness and distinct layers, with a penetration depth of several hundred microns in tissue [20]. The backscattered light is captured using an interferometric setup to

reconstruct the depth profile of the sample at the selected location. By acquiring a series of cross-sectional slices (B-scans), OCT produces high-resolution 3D images of the retina quickly, non-invasively, and painlessly. [77], [72]. OCT is similar to ultrasound imaging, but it uses light instead of sound. Developed in the 1990s [77], OCT became commercially available in 2006.

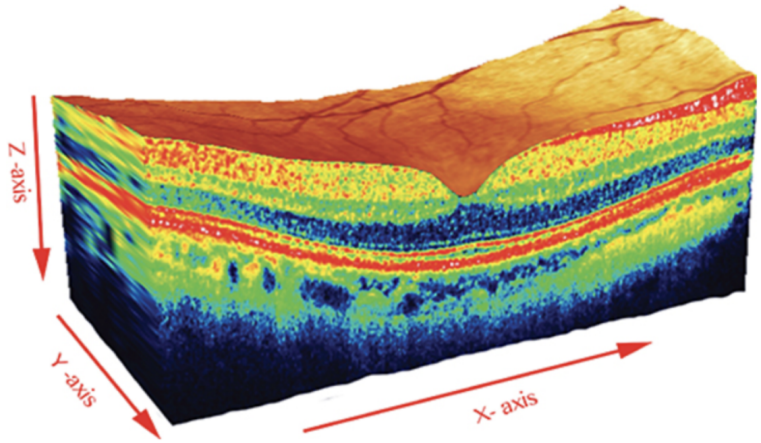

Figure A.5: An example of the retina Optical coherence tomography (OCT) volume.

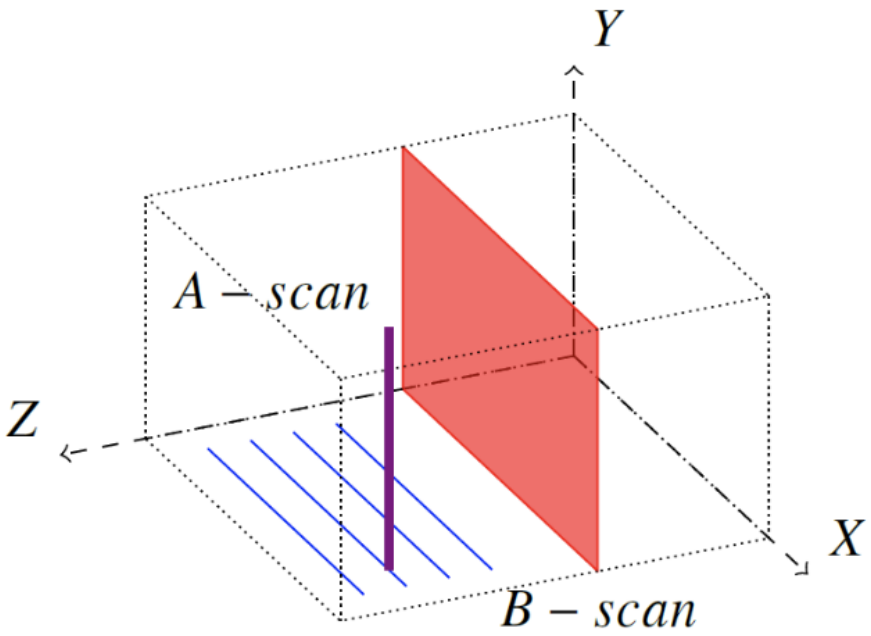

Figure A.6: OCT acquisition and the coordinate system: 1D axial scans (A-scans, purple) are combined to create 2D cross-sectional slices (B-scans, red) by scanning through the volume in a raster scan pattern (blue). Multiple B-scans are then compiled to form a complete OCT volume. Image taken from [21]

The OCT imaging technique involves multiple steps and a specific coordinate system, as illustrated in Figures A.6 and A.7 and briefly discussed:
A-scan (Axial Scan): The fundamental unit of OCT imaging is the A-scan, a one-dimensional depth scan that measures the reflection of light at different depths in

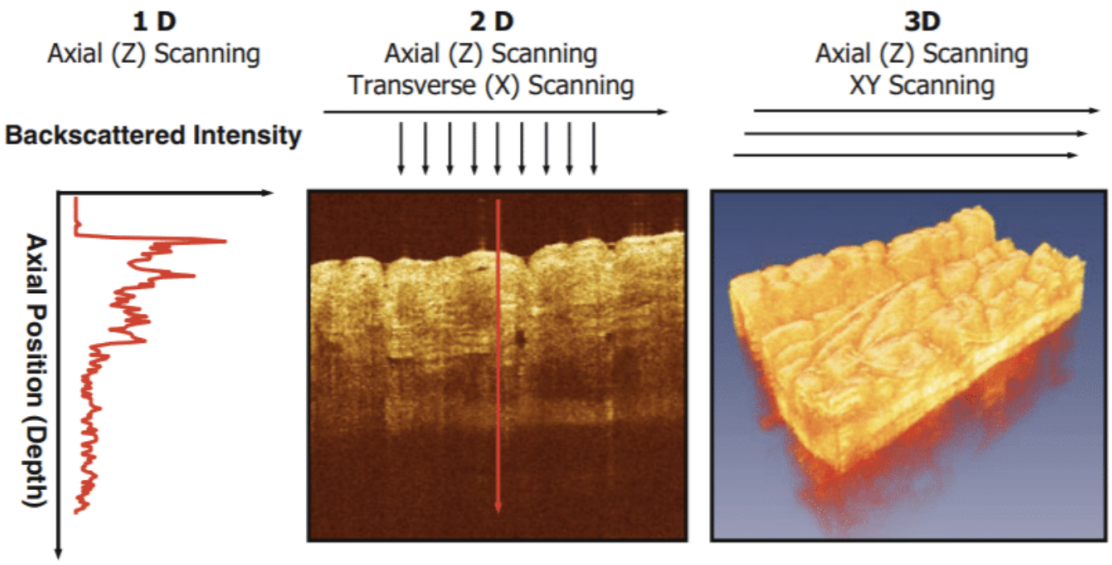

Figure A.7: An example illustration of a retina Optical Coherence Tomography (OCT) image showing an A-scan, B-scan, and 3D views. Image taken from [162]

the retina. Each A-scan captures the reflectivity profile along a single line through the tissue, providing information about the internal structure of the retina along that line. In diagrams, A-scans are often represented in purple.

B-scan (Cross-Sectional Scan): Multiple A-scans are combined side-by-side to create a two-dimensional cross-sectional image, known as a B-scan. This 2D slice provides a detailed view of the retina's layers and any abnormalities present within that plane. B-scans are typically represented in red. The B-scan is generated by moving the OCT beam across the retina in a linear fashion, collecting a series of adjacent A-scans.

Raster Scan Pattern: To form a three-dimensional (3D) volume, the OCT system performs multiple B-scans in a systematic manner, often following a raster scan pattern. In this pattern, the OCT beam is moved in a grid-like fashion across the retinal surface, capturing a series of B-scans that cover the entire area of interest. This process is shown in blue in illustrations.

OCT Volume: The complete OCT volume is constructed by stacking these sequential B-scans together. This 3D representation allows for the examination of the retinal structure in great detail, providing valuable insights into various retinal conditions and diseases.

The coordinate system for OCT imaging typically includes:

X-axis: Represents the lateral (horizontal) dimension within the plane of the retina, corresponding to the direction of the B-scan.

Y-axis: Represents the axial (depth) dimension, corresponding to the direction of the A-scan.

Z-axis: Represents the second lateral (vertical) dimension, perpendicular to the direction of the B-scan, and it corresponds to the stacking of multiple B-scans to create the volume.

The detailed 3D visualization provided by OCT allows for early detection and monitoring of diseases such as Age-related Macular Degeneration (AMD) and Diabetic Macular Edema (DME). An illustration of the OCT scan is shown in Figure A.5.

OCT Device Manufacturers

Optical Coherence Tomography (OCT) devices are manufactured by several companies, each offering unique technological features and specifications tailored for ophthalmic imaging. Here's an overview of some prominent OCT device manufacturers:

Carl Zeiss Meditec
Device Series: Cirrus OCT
Features: Cirrus OCT systems are known for their high-resolution imaging capabilities, offering detailed visualization of retinal layers and structures. They use spectral domain OCT (SD-OCT) technology, which provides faster image acquisition and higher axial resolution compared to earlier time domain OCT (TD-OCT) systems.

Heidelberg Engineering
Device Series: Spectralis OCT
Features: Spectralis OCT systems combine OCT with confocal scanning laser ophthalmoscopy (cSLO). This integration allows simultaneous imaging of retinal structures and fundus autofluorescence (FAF). Spectralis devices are noted for their depth-resolved imaging and tracking features, enabling precise alignment for follow-up scans over time.

Topcon
Device Series: Topcon OCT
Features: Topcon offers a range of OCT systems catering to different clinical needs. Their devices utilize spectral domain OCT (SD-OCT) technology, providing high-resolution cross-sectional images of the retina. Topcon OCT systems are recognized for their ease of use, advanced image processing, and comprehensive analysis software.
Each manufacturer may have multiple models within their OCT device series, varying in features such as scan speed, resolution, depth penetration, and additional functionalities like angiography or widefield imaging capabilities. These devices play a critical role in diagnosing and monitoring various retinal diseases, including diabetic retinopathy, age-related macular degeneration, and glaucoma, by providing detailed structural information of the retina non-invasively and in real-time.

Differences Between Retinal Fundus and OCT Images

Retinal fundus photography and retinal Optical Coherence Tomography (OCT) are two distinct imaging techniques used in ophthalmology for evaluating the retina, each offering unique perspectives and information. Here are some of the key differences and complementarities between these imaging techniques:

Nature of Imaging:
Fundus Photography: Provides a wide-angle, 2D view of the retina, including the optic disc, macula, and peripheral retina.
OCT: Produces high-resolution, 3D cross-sectional images of retinal layers and structures, offering detailed information about retinal thickness and morphology.

Information Provided:
Fundus Photography: Offers a panoramic view useful for identifying surface-level abnormalities like hemorrhages, drusen, and vascular changes.
OCT: Enables visualization of individual retinal layers, allowing detection of subtle changes such as fluid accumulation, retinal thinning, or thickening.

Clinical Use:
Fundus Photography: Commonly used for screening, monitoring diabetic retinopathy, hypertensive retinopathy, and macular degeneration.
OCT: Essential for diagnosing and managing diseases affecting retinal structure, including macular edema, retinal detachment, and glaucoma.

Resolution and Depth:
Fundus Photography: Lower resolution compared to OCT, but captures a wide field of view in a single image.
OCT: Higher resolution and depth-resolved images, providing quantitative data on retinal thickness and pathology.

Complementarity:
Fundus Photography: Provides context and overview of retinal health, guiding the need for further OCT examination.
OCT: Offers detailed structural information not visible with fundus photography alone, supporting precise diagnosis and treatment monitoring.

In clinical practice, these techniques are often used complementarily: fundus photography for initial screening and broad assessment, and OCT for detailed evaluation of specific retinal structures and pathology. Together, they provide comprehensive insights into retinal health and pathology, aiding in the management of various eye diseases.

A.1.1 Head and Neck Cancer Overview

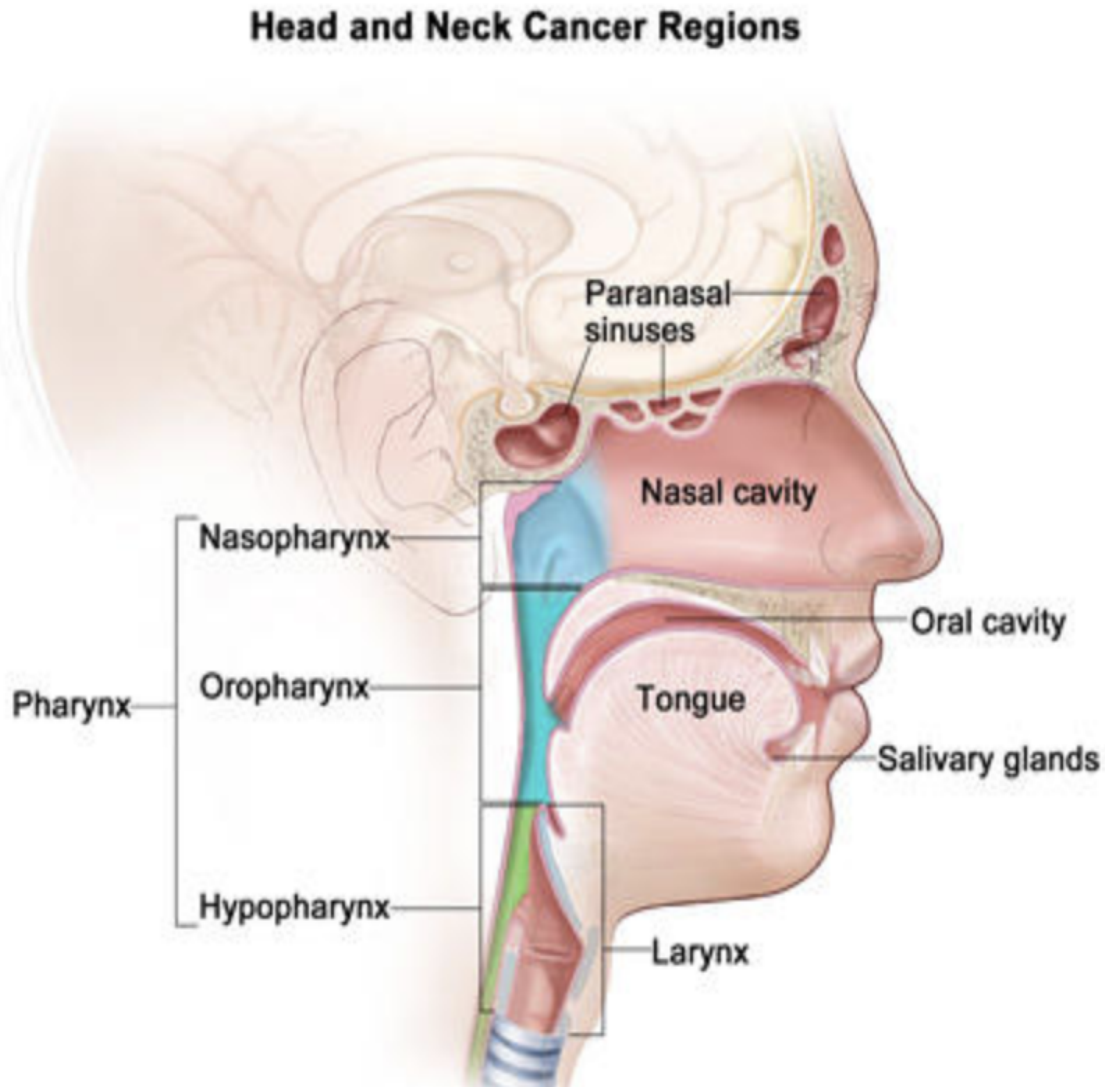

Figure A.8: An illustration showing the head and neck cancer regions. Image taken from [70]

Head and neck cancer is a broad term encompassing various cancers in the larynx, throat, lips, mouth, nose, and salivary glands. These cancers typically originate in the squamous cells lining the mucosal surfaces of these areas, such as the inside of the mouth, throat, and voice box, and are known as squamous cell carcinomas of the head and neck. Although it is less common, head and neck cancer can also begin in the salivary glands, sinuses, muscles, or nerves in these regions. Head and neck cancers are among the most common cancers globally, ranking as the 5th leading cancer by incidence [41], [175], [47]. Some of the most common forms of head and neck cancers include:

Oral cancer: Forms in the lips, tongue, gums, the lining of the cheeks and lips, the roof and floor of the mouth, or behind the wisdom teeth.

Oropharyngeal cancer: Affects the middle part of the throat (oropharynx), with tonsil cancer being the most common type.

Hypopharyngeal cancer: Affects the bottom part of the throat (hypopharynx).

Laryngeal cancer: Affects the voice box (larynx), which houses the vocal cords.

Nasopharyngeal cancer: Affects the upper part of the throat (nasopharynx).

Salivary gland cancer: Affects the salivary glands, which produce saliva.

Nasal cavity and paranasal sinus cancer: Forms in the hollow area inside the nose (nasal cavity) or the hollow spaces in the bones surrounding the nose (paranasal sinuses).

Head and neck cancer sometimes spread to the lymph nodes in the upper part of the neck. Despite their locations, brain, eye, esophageal, and thyroid cancers aren't typically considered head and neck cancers, as they require different treatments from those used for head and neck cancer. An illustration of the head and neck cancer regions is shown in Figure A.8

Symptoms of Head and Neck Cancer

Figure A.9: A diagram summarizing the symptoms of head and neck cancer.

The symptoms of head and neck cancer vary depending on the location where the cancer originates. Some of the most common symptoms include:
Lumps or Swelling: A persistent lump or swelling in the neck, throat, or jaw that does not go away. A lump or sore inside the mouth.
Sore Throat: A persistent sore throat or the feeling that something is caught in the throat. Pain or difficulty swallowing (dysphagia).
Voice Changes: Hoarseness or changes in the voice that last more than two weeks. Persistent cough or coughing up blood.
Mouth Issues: Red or white patches in the mouth or on the tongue. Unexplained bleeding in the mouth. Persistent sores or ulcers in the mouth that do not heal.
Nasal Symptoms: Nasal congestion or blocked sinuses that do not clear. Frequent nosebleeds or unusual nasal discharge. Pain around the upper teeth or problems with dentures.
Ear Pain: Ear pain or trouble hearing. Ringing in the ears (tinnitus).

Weight Loss: Unexplained weight loss.
Difficulty Breathing: Shortness of breath or noisy breathing.
Facial Pain or Numbness: Persistent pain or numbness in the face or neck.
Difficulty Moving the Jaw: Problems with jaw movement or pain when opening the mouth. A summary of the symptoms of head and neck cancer is illustrated in Figure A.9

A.2 Treatment for Head and Neck Cancer

Effective treatments for head and neck cancer include surgery, radiation therapy, and chemotherapy, either individually or in combination. The specific treatment plan depends on various factors including the cancer's stage and location. Early detection and ongoing monitoring are crucial for successful treatment outcomes. These methods will be briefly discussed.

Surgery: There are three main types of surgery used to treat head and neck cancer,
Tumor Resection: Removal of the tumor and some surrounding healthy tissue.
Neck Dissection: Removal of lymph nodes if the cancer has spread.
Reconstructive Surgery: To restore function and appearance after tumor removal.

Radiation Therapy: High-energy beams are used to destroy cancer cells. Can be used alone or in combination with surgery and/or chemotherapy. Types include external beam radiation and brachytherapy.

Chemotherapy: Uses drugs to kill cancer cells or stop them from growing. Often combined with radiation therapy (chemoradiation) for greater effectiveness. The two main types of chemotherapy are:
Targeted Therapy: Drugs designed to target specific molecules involved in cancer cell growth. Examples include monoclonal antibodies and tyrosine kinase inhibitors.
Immunotherapy: Boosts the body's immune system to fight cancer. Checkpoint inhibitors are a common type used in head and neck cancer.

Head and Neck Cancer Imaging Techniques

Imaging techniques for head and neck cancers are crucial for diagnosis, staging, treatment planning, and monitoring response to therapy. Some of the primary imaging techniques include:

Computed Tomography (CT) Scan: Provides detailed cross-sectional images of the head and neck region, aiding in determining the size, shape, and location of tumors. It is useful for detecting lymph node involvement and assessing the extent of the disease.

Magnetic Resonance Imaging (MRI): Offers high-resolution images of soft tissues, making it superior for evaluating the extent of tumor invasion into surrounding tissues such as muscles, nerves, and blood vessels. It is particularly useful for cancers in complex anatomical areas like the base of the skull.

Positron Emission Tomography (PET) Scan: Often combined with CT (PET/CT) to provide both metabolic and anatomical information. It helps detect metastases, evaluate the metabolic activity of the tumor, and is useful for staging and assessing treatment response.

Ultrasound: Used to evaluate cervical lymph nodes and guide fine-needle aspiration biopsies. It is non-invasive and readily available.

X-rays: Generally used for initial assessment and in conjunction with other imaging modalities. Panoramic dental X-rays can help detect oral cancers involving the jaw.

Endoscopy: Involves inserting a flexible tube with a camera (endoscope) through the nose or mouth to visualize internal structures. It is useful for direct visualization and biopsy of tumors in the upper aerodigestive tract.

Barium Swallow: A specialized X-ray procedure where the patient swallows a barium-containing liquid. It helps visualize the esophagus and detect abnormalities

related to swallowing issues.
Sialography: An imaging technique used to evaluate the salivary glands. It involves injecting a contrast agent into the salivary ducts followed by X-ray imaging. These imaging techniques provide critical information that aids in the comprehensive management of head and neck cancers, from initial diagnosis to treatment planning and follow-up.

Publications

1. Nchongmaje Ndipenoch et al. “Simultaneous segmentation of layers and fluids in retinal oct images”. In: 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE. 2022, pp. 1–6.

2. McConnell, N., Ndipenoch, N., Cao, Y., Miron, A., Li, Y “Exploring advanced architectural variations of nnUNet”. In: Neurocomputing 560 (2023), p. 126837.

3. Nchongmaje Ndipenoch, Alina Miron, and Yongmin Li. “Performance Evaluation of Retinal OCT Fluid Segmentation, Detection, and Generalization Over Variations of Data Sources”. In: IEEE Access 12 (2024), pp. 3171931735.

4. Nchongmaje Ndipenoch et al. “CVD Net: Head and Neck Tumor Segmentation and Generalization in PET/CT Scans Across Data from Multiple Medical Centers”. In: 2024 1st International Conference on AI in Healthcare. Springer. 2024, pp. 64–76.

5. Luo, X., Fu, J., Zhong, Y., Liu, S., Han, B., Astaraki, N. Ndipenoch, and Zhang, S. (2025)... Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. Medical image analysis, 101, 103447.

6. MMIS-Net for Retinal Fluid Segmentation and Detection **(Under Review)**

Bibliography

[1] Michael D Abràmoff, Mona K Garvin, and Milan Sonka. "Retinal imaging and image analysis". In: *IEEE reviews in biomedical engineering* 3 (2010), pp. 169–208.

[2] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[3] Amina Adadi. "A survey on data-efficient algorithms in big data era". In: *Journal of Big Data* 8.1 (2021), p. 24.

[4] Ahmed Almazroa et al. "Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images". In: *International ophthalmology* 37 (2017), pp. 701–717.

[5] Mina Amiri, Rupert Brooks, and Hassan Rivaz. "Fine-tuning U-Net for ultrasound image segmentation: different layers, different outcomes". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67.12 (2020), pp. 2510–2518.

[6] Chengyang An, Huai Chen, and Lisheng Wang. "A coarse-to-fine framework for head and neck tumor segmentation in CT and PET images". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2021, pp. 50–57.

[7] Vincent Andrearczyk et al. "Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images". In: *3D head and neck tumor segmentation in PET/CT challenge*. Springer, 2021, pp. 1–37.

[8] Michela Antonelli et al. "The medical segmentation decathlon". In: *Nature communications* 13.1 (2022), p. 4128.

[9] Antreas Antoniou, Harri Edwards, and Amos Storkey. "How to train your MAML". In: *Seventh International Conference on Learning Representations*. 2019.

[10] Stefanos Apostolopoulos et al. "Simultaneous classification and segmentation of cysts in retinal OCT". In: *Proc. MICCAI Retinal OCT Fluid Challenge (RETOUCH)*. 2017, pp. 22–29.

[11] *Automated Cardiac Diagnosis Challenge (ACDC)*. https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html.

[12] Ujjwal Baid et al. "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv 2021". In: *arXiv preprint arXiv:2107.02314* (2021).

[13] Spyridon Bakas et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13.

[14] Spyridon Bakas et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (2018).

[15] Tatiana Bejarano, Mariluz Ornelas-Couto, and Ivaylo Mihaylov. "Head-and-neck squamous cell carcinoma (HNSCC) patients with 3D CT taken during pre-treatment, mid-treatment, and post-treatment Dataset". In: (2018).

[16] Shai Ben-David et al. "Analysis of representations for domain adaptation". In: *Advances in neural information processing systems* 19 (2006).

[17] Qi Bi et al. "Learning Generalized Medical Image Segmentation from Decoupled Feature Queries". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2. 2024, pp. 810–818.

[18] Cheng Bian et al. "Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation". In: *IEEE Transactions on Medical Imaging* 41.5 (2021), pp. 1043–1056.

[19] Patrick Bilic et al. "The liver tumor segmentation benchmark (lits)". In: *Medical Image Analysis* 84 (2023), p. 102680.

[20] Josef F Bille. "High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics". In: (2019).

[21] Hrvoje Bogunović et al. "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge". In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1858–1874.

[22] George H Bresnick. "Diabetic macular edema: a review". In: *Ophthalmology* 93.7 (1986), pp. 989–997.

[23] David Brown et al. "Current best clinical practices—management of neovascular AMD". In: *Journal of vitreoretinal diseases* 1.5 (2017), pp. 294–297.

[24] *Cancer Imaging Archiv*. https://www.cancerimagingarchive.net/.

[25] J Quinonero Candela et al. "Dataset shift in machine learning". In: *The MIT Press* 1 (2009), p. 5.

[26] Aaron Carass et al. "Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis". In: *Scientific reports* 10.1 (2020), p. 8242.

[27] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[28] Qi Chang et al. "Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 13856–13866.

[29] Woong-Gi Chang et al. "Domain-specific batch normalization for unsupervised domain adaptation". In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2019, pp. 7354–7362.

[30] Yao Chang et al. "Transclaw u-net: Claw u-net with transformers for medical image segmentation". In: *arXiv preprint arXiv:2107.05188* (2021).

[31] Cheng Chen et al. "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation". In: *Medical Image Analysis* (2024), p. 103310.

[32] Cheng Chen et al. "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 865–872.

[33] Jieneng Chen et al. "3d transunet: Advancing medical image segmentation through vision transformers". In: *arXiv preprint arXiv:2310.07781* (2023).

[34] Jieneng Chen et al. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[35] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[36] Sihong Chen, Kai Ma, and Yefeng Zheng. "Med3d: Transfer learning for 3d medical image analysis". In: *arXiv preprint arXiv:1904.00625* (2019).

[37] Xiaoyang Chen et al. "Versatile medical image segmentation learned from multi-source datasets via model self-disambiguation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11747–11756.

[38] Xu Chen et al. "Diverse data augmentation for learning image segmentation with cross-modality annotations". In: *Medical image analysis* 71 (2021), p. 102060.

[39] Dongjie Cheng et al. "Sam on medical images: A comprehensive study on three prompt modes". In: *arXiv preprint arXiv:2305.00035* (2023).

[40] Stephanie J Chiu et al. "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema". In: *Biomedical optics express* 6.4 (2015), pp. 1172–1194.

[41] Laura QM Chow. "Head and neck cancer". In: *New England Journal of Medicine* 382.1 (2020), pp. 60–72.

[42] Jaime A Davidson et al. "How the diabetic eye loses vision". In: *Endocrine* 32 (2007), pp. 107–116.

[43] Jeffrey De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". In: *Nature medicine* 24.9 (2018), pp. 1342–1350.

[44] Ruining Deng et al. "Omni-seg: A single dynamic network for multi-label renal pathology image segmentation using partially labeled data". In: *arXiv preprint arXiv:2112.12665* (2021).

[45] Foivos I Diakogiannis et al. "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020), pp. 94–114.

[46] P Kingma Diederik. "Adam: A method for stochastic optimization". In: *(No Title)* (2014).

[47] Parkin Dm. "Global cancer statistics, 2002". In: *CA Cancer J Clin* 55 (2005), pp. 74–108.

[48] Konstantin Dmitriev and Arie E Kaufman. "Learning multi-class segmentations from single-class datasets". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9501–9511.

[49] Bashir I Dodo et al. "Level Set Segmentation of Retinal OCT Images." In: *BIOIMAGING*. 2019, pp. 49–56.

[50] Bashir Isa Dodo et al. "Retinal layer segmentation in optical coherence tomography images". In: *IEEE Access* 7 (2019), pp. 152388–152398.

[51] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[52] Qi Dou et al. "Unpaired multi-modal segmentation via knowledge distillation". In: *IEEE transactions on medical imaging* 39.7 (2020), pp. 2415–2425.

[53] Siyi Du et al. "Mdvit: Multi-domain vision transformer for small medical image segmentation datasets". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 448–458.

[54] Leyuan Fang et al. "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search". In: *Biomedical optics express* 8.5 (2017), pp. 2732–2744.

[55] Xi Fang and Pingkun Yan. "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction". In: *IEEE Transactions on Medical Imaging* 39.11 (2020), pp. 3619–3629.

[56] Botond Fazekas et al. "Adapting Segment Anything Model (SAM) for Retinal OCT". In: *International Workshop on Ophthalmic Medical Image Analysis*. Springer. 2023, pp. 92–101.

[57] Weijia Feng, Lingting Zhu, and Lequan Yu. "Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars". In: *arXiv preprint arXiv:2308.14133* (2023).

[58] Delia Cabrera Fernandez. "Delineating fluid-filled region boundaries in optical coherence tomography images of the retina". In: *IEEE transactions on medical imaging* 24.8 (2005), pp. 929–945.

[59] Robert M French. "Catastrophic forgetting in connectionist networks". In: *Trends in cognitive sciences* 3.4 (1999), pp. 128–135.

[60] Yunhe Gao. "Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11194–11204.

[61] Yunhe Gao et al. "A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark". In: *arXiv preprint arXiv:2203.00131* (2022).

[62] Mona Kathryn Garvin et al. "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images". In: *IEEE transactions on medical imaging* 28.9 (2009), pp. 1436–1447.

[63] Sergios Gatidis et al. "A whole-body FDG-PET/CT dataset with manually annotated tumor lesions". In: *Scientific Data* 9.1 (2022), p. 601.

[64] Ruiquan Ge et al. "MD-UNET: Multi-input dilated U-shape neural network for segmentation of bladder cancer". In: *Computational Biology and Chemistry* 93 (2021), p. 107510.

[65] Hao Guan and Mingxia Liu. "Domain adaptation for medical image analysis: a survey". In: *IEEE Transactions on Biomedical Engineering* 69.3 (2021), pp. 1173–1185.

[66] Ali Hatamizadeh et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images". In: *International MICCAI brainlesion workshop*. Springer. 2021, pp. 272–284.

[67] Ali Hatamizadeh et al. "Unetr: Transformers for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 574–584.

[68] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[69] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[70] *Head and neck Cancer Regions*. https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet.

[71] Nicholas Heller et al. "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes". In: *arXiv preprint arXiv:1904.00445* (2019).

[72] Christoph K HITZENBERGER. "Optical Measurement of the Axial Eye Length by Laser Doppler Interferometry". In: *SPIE milestone series* 165 (2001), pp. 260–268.

[73] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

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

[75] Shishuai Hu et al. "Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation". In: *IEEE Transactions on Medical Imaging* 42.1 (2022), pp. 233–244.

[76] Chao Huang et al. "3D U 2-Net: A 3D universal U-Net for multi-domain medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 291–299.

[77] David Huang et al. "Optical coherence tomography". In: *science* 254.5035 (1991), pp. 1178–1181.

[78] Lina Huang et al. "Segmenting Medical Images: From UNet to Res-UNet and nnUNet". In: *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2024, pp. 483–489.

[79] Rui Huang et al. "Multi-organ segmentation via co-training weight-averaged models from few-organ datasets". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer. 2020, pp. 146–155.

[80] Zhongmiao Huang, Liejun Wang, and Lianghui Xu. "DRA-Net: Medical image segmentation based on adaptive feature extraction and region-level information fusion". In: *Scientific Reports* 14.1 (2024), p. 9714.

[81] *Human Eye Anatomy*. https://www.elmanretina.com/the-basic-anatomy-of-the-retina/.

[82] Andrei Iantsen, Dimitris Visvikis, and Mathieu Hatt. "Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images". In: *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*. Springer. 2021, pp. 37–43.

[83] Fabian Isensee et al. "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 488–498.

[84] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[85] Kazufumi Ito and Kaiqi Xiong. "Gaussian filters for nonlinear filtering problems". In: *IEEE transactions on automatic control* 45.5 (2000), pp. 910–927.

[86] C Jacobs et al. *LUNA-16: Lung Nodule Analysis*. 2017.

[87] Debesh Jha et al. "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation". In: *IEEE journal of biomedical and health informatics* 25.6 (2021), pp. 2029–2040.

[88] Debesh Jha et al. "Resunet++: An advanced architecture for medical image segmentation". In: *2019 IEEE international symposium on multimedia (ISM)*. IEEE. 2019, pp. 225–2255.

[89] Yuanfeng Ji et al. "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation". In: *Advances in neural information processing systems* 35 (2022), pp. 36722–36732.

[90] Zhanghexuan Ji et al. "Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21140–21151.

[91] Yun Jiang et al. "SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer". In: *Brain sciences* 12.6 (2022), p. 797.

[92] C Daniel Johnson et al. "Accuracy of CT colonography for detection of large adenomas and cancers". In: *New England Journal of Medicine* 359.12 (2008), pp. 1207–1217.

[93] Sung Ho Kang et al. "Deep neural networks for the detection and segmentation of the retinal fluid in OCT images". In: *MICCAI Retinal OCT Fluid Challenge (RETOUCH)* (2017).

[94] Neerav Karani et al. "A lifelong learning approach to brain MR segmentation across scanners and protocols". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 476–484.

[95] A Emre Kavur et al. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation". In: *Medical Image Analysis* 69 (2021), p. 101950.

[96] Diederik P Kingma. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[97] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[98] *kits19 Dataset*. https://kits19.grand-challenge.org/data/.

[99] Valentin Koch et al. "Noise transfer for unsupervised domain adaptation of retinal OCT images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 699–708.

[100] Bennett Landman et al. "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge". In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. Vol. 5. 2015, p. 12.

[101] Cecilia S Lee et al. "Deep-learning based, automated segmentation of macular edema in optical coherence tomography". In: *Biomedical optics express* 8.7 (2017), pp. 3440–3448.

[102] Wenhui Lei et al. "Contrastive learning of relative position regression for one-shot object localization in 3D medical images". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer. 2021, pp. 155–165.

[103] Wenhui Lei et al. "Medlsam: Localize and segment anything model for 3d medical images". In: *arXiv preprint arXiv:2306.14752* (2023).

[104] Tianang Leng et al. "Self-sampling meta SAM: enhancing few-shot medical image segmentation with meta-learning". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7925–7935.

[105] Daiqing Li et al. "Federated simulation for medical imaging". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer. 2020, pp. 159–168.

[106] Jeany Q Li et al. "Retinal diseases in Europe". In: *European Society of Retina Specialists (EURETINA)* (2017).

[107] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks". In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 4313–4324.

[108] Laurence S Lim et al. "Age-related macular degeneration". In: *The Lancet* 379.9827 (2012), pp. 1728–1738.

[109] Manying Lin, Qingling Cai, and Jun Zhou. "3D Md-Unet: A novel model of multi-dataset collaboration for medical image segmentation". In: *Neurocomputing* 492 (2022), pp. 530–544.

[110] Jie Liu et al. "Clip-driven universal model for organ segmentation and tumor detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21152–21164.

[111] Pengbo Liu et al. "Universal segmentation of 33 anatomies". In: *arXiv preprint arXiv:2203.02098* (2022).

[112] Quande Liu et al. "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1013–1023.

[113] Quande Liu et al. "MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data". In: *IEEE transactions on medical imaging* 39.9 (2020), pp. 2713–2724.

[114] Shikun Liu, Edward Johns, and Andrew J Davison. "End-to-end multi-task learning with attention". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1871–1880.

[115] Siqi Liu et al. "3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer. 2018, pp. 851–858.

[116] Wentao Liu et al. "LiteMedSAM with Low-Rank Adaptation and Multi-Box Efficient Inference for Medical Image Segmentation". In: ().

[117] Ziwei Liu et al. "Large-scale long-tailed recognition in an open world". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2537–2546.

[118] Donghuan Lu et al. "Retinal fluid segmentation and detection in optical coherence tomography images using fully convolutional neural network". In: *arXiv preprint arXiv:1710.04778* (2017).

[119] Jiangshan Lu et al. "Priori and posteriori attention for generalizing head and neck tumors segmentation". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2021, pp. 134–140.

[120] Xiangde Luo et al. "Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma". In: *arXiv preprint arXiv:2312.09576* (2023).

[121] Jun Ma, Feifei Li, and Bo Wang. "U-mamba: Enhancing long-range dependency for biomedical image segmentation". In: *arXiv preprint arXiv:2401.04722* (2024).

[122] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654.

[123] Matthis Manthe, Stefan Duffner, and Carole Lartizien. "Federated brain tumor segmentation: an extensive benchmark". In: *Medical Image Analysis* 97 (2024), p. 103270.

[124] Matthis Manthe, Stefan Duffner, and Carole Lartizien. "Whole brain radiomics for clustered federated personalization in brain tumor segmentation". In: *Medical Imaging with Deep Learning*. PMLR. 2024, pp. 957–977.

[125] Maciej A Mazurowski et al. "Segment anything model for medical image analysis: an experimental study". In: *Medical Image Analysis* 89 (2023), p. 102918.

[126] Niccolo McConnell et al. "Integrating residual, dense, and inception blocks into the nnunet". In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2022, pp. 217–222.

[127] Niccolò McConnell et al. "Exploring advanced architectural variations of nnUNet". In: *Neurocomputing* 560 (2023), p. 126837.

[128] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[129] *Medical Decathlon Dataset*. http://medicaldecathlon.com/.

[130] Martina Melinščak et al. "Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation". In: *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije* 62.3-4 (2021), pp. 375–385.

[131] Martina Melinščak et al. "Aroi: Annotated retinal oct images database". In: *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE. 2021, pp. 371–376.

[132] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[133] Andriy Myronenko et al. "Automated head and neck tumor segmentation from 3D PET/CT HECKTOR 2022 challenge report". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2022, pp. 31–37.

[134] Ying-Hwey Nai et al. "Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset". In: *Computers in biology and medicine* 134 (2021), p. 104497.

[135] Mohamed A Naser et al. "Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in PET/CT images". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2021, pp. 121–133.

[136] Nchongmaje Ndipenoch, Alina Miron, and Yongmin Li. "Performance Evaluation of Retinal OCT Fluid Segmentation, Detection, and Generalization Over Variations of Data Sources". In: *IEEE Access* 12 (2024), pp. 31719–31735.

[137] Nchongmaje Ndipenoch et al. "CVD_Net: Head and Neck Tumor Segmentation and Generalization in PET/CT Scans Across Data from Multiple Medical Centers". In: *International Conference on AI in Healthcare*. Springer. 2024, pp. 64–76.

[138] Nchongmaje Ndipenoch et al. "nnUNet RASPP for Retinal OCT Fluid Detection, Segmentation and Generalisation over Variations of Data Sources". In: *arXiv preprint arXiv:2302.13195* (2023).

[139] Nchongmaje Ndipenoch et al. "Retinal image segmentation with small datasets". In: *arXiv preprint arXiv:2303.05110* (2023).

[140] Nchongmaje Ndipenoch et al. "Simultaneous segmentation of layers and fluids in retinal oct images". In: *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2022, pp. 1–6.

[141] John A Onofrey et al. "Generalizable multi-site training and testing of deep neural networks using image normalization". In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 348–351.

[142] *Pancreas-CT Dataset*. https://www.cancerimagingarchive.net/collection/pancreas-ct/.

[143] Nishtha Panwar et al. "Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare". In: *Telemedicine and e-Health* 22.3 (2016), pp. 198–208.

[144] Jay N Paranjape et al. "Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation". In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2024, pp. 187–201.

[145] D Max Parkin et al. "Global cancer statistics, 2002". In: *CA: a cancer journal for clinicians* 55.2 (2005), pp. 74–108.

[146] Kelly Payette et al. "Multi-Center Fetal Brain Tissue Annotation (FeTA) Challenge 2022 Results". In: *arXiv preprint arXiv:2402.09463* (2024).

[147] Chenhao Pei et al. "Multi-source domain adaptation for medical image segmentation". In: *IEEE Transactions on Medical Imaging* (2023).

[148] Chao Peng et al. "Large kernel matters–improve semantic segmentation by global convolutional network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4353–4361.

[149] Shehan Perera, Pouyan Navard, and Alper Yilmaz. "SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 4981–4988.

[150] Alexander Pfefferle, Lennart Purucker, and Frank Hutter. "DAFT: Data-Aware Fine-Tuning of Foundation Models for Efficient and Effective Medical Image Segmentation". In: *CVPR 2024: Segment Anything In Medical Images On Laptop*. 2024.

[151] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[152] Abdolreza Rashno, Dara D Koozekanani, and Keshab K Parhi. "Detection and segmentation of various types of fluids with graph shortest path and deep learning approaches". In: *Proc. MICCAI Retinal OCT Fluid Challenge (RETOUCH)* (2017), pp. 54–62.

[153] Jintao Ren et al. "PET normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2021, pp. 83–91.

[154] *Retina Disease*. https://www.scienceofamd.org/.

[155] *Retina structure*. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/retina.

[156] *Retouch Grand*. https://retouch.grand-challenge.org/Home/.

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

[158] T-YLPG Ross and GKHP Dollár. "Focal loss for dense object detection". In: *proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2980–2988.

[159] Holger R Roth et al. "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer. 2015, pp. 556–564.

[160] Abhijit Guha Roy et al. "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks". In: *Biomedical optics express* 8.8 (2017), pp. 3627–3642.

[161] Leonardo Rundo et al. "USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets". In: *Neurocomputing* 365 (2019), pp. 31–43.

[162] Sunčica Sakić. "Modelling Light Propagation In Ocular Tissues". PhD thesis. Sept. 2020.

[163] Ana Salazar-Gonzalez, Yongmin Li, and Djibril Kaba. "MRF reconstruction of retinal images for the optic disc segmentation". In: *Health Information Science: First International Conference, HIS 2012, Beijing, China, April 8-10, 2012. Proceedings 1*. Springer. 2012, pp. 88–99.

[164] Ana Salazar-Gonzalez et al. "Segmentation of the blood vessels and optic disk in retinal images". In: *IEEE journal of biomedical and health informatics* 18.6 (2014), pp. 1874–1886.

[165] Ana G Salazar-Gonzalez, Yongmin Li, and Xiaohui Liu. "Retinal blood vessel segmentation via graph cut". In: *2010 11th International Conference on Control Automation Robotics & Vision*. IEEE. 2010, pp. 225–230.

[166] Adam Santoro et al. "Meta-learning with memory-augmented neural networks". In: *International conference on machine learning*. PMLR. 2016, pp. 1842–1850.

[167] Sourya Sengupta, Satrajit Chakrabarty, and Ravi Soni. "Is SAM 2 Better than SAM in Medical Image Segmentation?" In: *arXiv preprint arXiv:2408.04212* (2024).

[168] Micah J Sheller et al. "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer. 2019, pp. 92–104.

[169] Chen Shen et al. "Multi-task federated learning for heterogeneous pancreas segmentation". In: *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2*. Springer. 2021, pp. 101–110.

[170] Chuyun Shen et al. "Interactive 3d medical image segmentation with sam 2". In: *arXiv preprint arXiv:2408.02635* (2024).

[171] Gonglei Shi et al. "Marginal loss and exclusion loss for partially supervised multi-organ segmentation". In: *Medical Image Analysis* 70 (2021), p. 101979.

[172] Gary Shienbaum et al. "Management of submacular hemorrhage secondary to neovascular age-related macular degeneration with anti–vascular endothelial growth factor monotherapy". In: *American journal of ophthalmology* 155.6 (2013), pp. 1009–1013.

[173] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. "Cancer statistics, 2018". In: *CA: a cancer journal for clinicians* 68.1 (2018), pp. 7–30.

[174] Santiago Silva et al. "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data". In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 270–274.

[175] Eugene Son et al. "Cancers of the major salivary gland". In: *Journal of oncology practice* 14.2 (2018), pp. 99–108.

[176] *structseg2019 Dataset*. https://structseg2019.grand-challenge.org/Home/.

[177] *synapse Dataset*. https://www.synapse.org/Synapse:syn3193805/wiki/89480.

[178] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC medical imaging* 15 (2015), pp. 1–28.

[179] Nima Tajbakhsh et al. "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.

[180] Ruwan Tennakoon et al. "Retinal fluid segmentation in OCT images using adversarial loss based convolutional neural networks". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1436–1440.

[181] Catalina Tobon-Gomez et al. "Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets". In: *IEEE transactions on medical imaging* 34.7 (2015), pp. 1460–1473.

[182] Antonio Torralba and Alexei A Efros. "Unbiased look at dataset bias". In: *CVPR 2011*. IEEE. 2011, pp. 1521–1528.

[183] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[184] Constantin Ulrich et al. "Multitalent: A multi-dataset approach to medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 648–658.

[185] LG Valiant. "A theory of the learnable‖ Communications of the ACM, 27 (11): 1134-1142". In: *Google Scholar Google Scholar Digital Library Digital Library* (1984).

[186] Vanya V Valindria et al. "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 547–556.

[187] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[188] Changyan Wang et al. "SAM-IE: SAM-based image enhancement for facilitating medical image diagnosis with segmentation foundation model". In: *Expert Systems with Applications* 249 (2024), p. 123795.

[189] Chuang Wang, Ya Xing Wang, and Yongmin Li. "Automatic choroidal layer segmentation using markov random field and level set method". In: *IEEE journal of biomedical and health informatics* 21.6 (2017), pp. 1694–1702.

[190] Guotai Wang et al. "Interactive medical image segmentation using deep learning with image-specific fine tuning". In: *IEEE transactions on medical imaging* 37.7 (2018), pp. 1562–1573.

[191] Liwei Wang, Yan Zhang, and Jufu Feng. "On the Euclidean distance of images". In: *IEEE transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1334–1339.

[192] Shujun Wang et al. "Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4237–4248.

[193] Jeffry Wicaksana et al. "FedMix: Mixed supervised federated learning for medical image segmentation". In: *IEEE Transactions on Medical Imaging* 42.7 (2022), pp. 1955–1968.

[194] Hao Wu, Shuchao Pang, and Arcot Sowmya. "Tgnet: A task-guided network architecture for multi-organ and tumour segmentation from partially labelled datasets". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[195] Junde Wu and Min Xu. "One-prompt to segment all medical images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11302–11312.

[196] Junde Wu et al. "Medical sam adapter: Adapting segment anything model for medical image segmentation". In: *arXiv preprint arXiv:2304.12620* (2023).

[197] Yingda Xia et al. "Auto-FedAvg: learnable federated averaging for multi-institutional medical image segmentation". In: *arXiv preprint arXiv:2104.10195* (2021).

[198] Enze Xie et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[199] Yutong Xie et al. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 171–180.

[200] Gang Xing et al. "Multi-scale pathological fluid segmentation in OCT with a novel curvature loss in convolutional neural network". In: *IEEE Transactions on Medical Imaging* 41.6 (2022), pp. 1547–1559.

[201] An Xu et al. "Closing the generalization gap of cross-silo federated medical image segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20866–20875.

[202] Qing Xu et al. "ESP-MedSAM: Efficient Self-Prompting SAM for Universal Domain-Generalized Medical Image Segmentation". In: *arXiv preprint arXiv:2407.14153* (2024).

[203] Z Xu. *Multi-atlas labeling beyond the cranial vault-workshop and challenge (2016)*. 2017.

[204] Govind Yadav, B Annappa, and DN Sachin. "Abdominal Multi-Organ Segmentation Using Federated Learning". In: *2024 IEEE Region 10 Symposium (TENSYMP)*. IEEE. 2024, pp. 1–7.

[205] Liu Yang et al. "Skymath: Technical report". In: *arXiv preprint arXiv:2310.16713* (2023).

[206] Jin Ye et al. "Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks". In: *arXiv preprint arXiv:2311.11969* (2023).

[207] Yiwen Ye et al. "Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 508–518.

[208] Fereshteh Yousefirizi et al. "Segmentation and risk score prediction of head and neck cancers in PET/CT volumes with 3D U-Net and cox proportional hazard neural networks". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2021, pp. 236–247.

[209] Jianpeng Zhang et al. "Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1195–1204.

[210] Kaidong Zhang and Dong Liu. "Customized segment anything model for medical image segmentation". In: *arXiv preprint arXiv:2304.13785* (2023).

[211] Ling Zhang et al. "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation". In: *IEEE transactions on medical imaging* 39.7 (2020), pp. 2531–2540.

[212] Penghao Zhang et al. "Domain adaptation for medical image segmentation: a meta-learning method". In: *Journal of Imaging* 7.2 (2021), p. 31.

[213] Wen Zhang et al. "A survey on negative transfer". In: *IEEE/CAA Journal of Automatica Sinica* 10.2 (2022), pp. 305–329.

[214] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. "Segment anything model for medical image segmentation: Current applications and future directions". In: *Computers in Biology and Medicine* (2024), p. 108238.

[215] Yizhe Zhang et al. "Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 129–139.

[216] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[217] Hengshuang Zhao et al. "Pyramid scene parsing network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[218] Xingchen Zhao et al. "Robust white matter hyperintensity segmentation on unseen domain". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1047–1051.

[219] Daquan Zhou et al. "Deepvit: Towards deeper vision transformer". In: *arXiv preprint arXiv:2103.11886* (2021).

[220] Hong-Yu Zhou et al. "nnformer: Interleaved transformer for volumetric segmentation". In: *arXiv preprint arXiv:2109.03201* (2021).

[221] Yuyin Zhou et al. "Prior-aware neural network for partially-supervised multi-organ segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 10672–10681.

[222] Zongwei Zhou et al. “Models genesis”. In: *Medical image analysis* 67 (2021), p. 101840.

[223] Xiahai Zhuang. “Multivariate mixture model for myocardial segmentation combining multi-source images”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.12 (2018), pp. 2933–2946.

[224] Xiahai Zhuang and Juan Shen. “Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI”. In: *Medical image analysis* 31 (2016), pp. 77–87.

[225] Xiahai Zhuang et al. “Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge”. In: *Medical image analysis* 58 (2019), p. 101537.