Ministère de l'enseignement Supérieur et de la recherche Scientifique

وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculté de Technologie

Département Informatique

جامعة باجي مختار – عنابـــة

كـلـيـة التكنـولـوجـيـا

قســم الاعلام الالـي

Thèse

Présentée pour obtenir le diplôme de

Doctorat Troisième Cycle

Filière : Informatique

Spécialité : Intelligence Artificielle

Par :

BOUKHAMLA Assia

Thème :

Techniques de l'Intelligence Artificielle pour l'aide à la décision médicale : Application aux maladies cardiaques

Thèse soutenue en 2025 devant le jury composé de :

N°	Nom et prénom	Grade	Etablissement	Qualité
01	FARAH Nadir	Prof.	Université Badji Mokhtar -Annaba	Président
02	AZIZI Nabiha	Prof.	Université Badji Mokhtar -Annaba	Rapporteur
03	BELHAOUARI Samir Brahim	MCA	Université Hamad Ben Khalifa-Qatar	Co-rapporteur
04	DJEBBAR Akila	MCA	Université Badji Mokhtar -Annaba	Examinateur
05	MAAZOUZI Faiz	MCA	Université Mohamed-Chérif Messaadia-Souk Ahras	Examinateur
06	BENMACHICHE Abdelmadjid	MCA	Université Chadli Bendjedid-El Taref	Examinateur

Ministry of Higher Education and Scientific Research
وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculty of Technology
Department of Computer Science

جامعة باجي مختار – عنابـــــــــة
كلية الهندســـــة
قســـــم الإعلام الآلي

Thesis

Submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy (PhD)

Specialization: Artificial Intelligence

Academic track: Computer Science

By:
Assia BOUKHAMLA

Title:

Artificial Intelligence Techniques for Medical Decision Support: Application to Cardiovascular Diseases

Thesis defended in 2025 in front of the committee:

N°	Full Name	Grade	Establishment	Quality
01	FARAH Nadir	Prof.	Badji Mokhtar-Annaba University	Chair
02	AZIZI Nabiha	Prof.	Badji Mokhtar-Annaba University	Supervisor
03	BELHAOUARI Samir Brahim	MCA	Hamad Bin Khalifa University-Qatar	Co-supervisor
04	DJEBBAR Akila	MCA	Badji Mokhtar-Annaba University	Examiner
05	MAAZOUZI Faiz	MCA	Mohamed-Chérif Messaadia-Souk Ahras University	Examiner
06	BENMACHICHE Abdelmadjid	MCA	Chadli Bendjedid-El Taref University	Examiner

Abstract

Cardiovascular disease (CVD) remains a leading cause of death worldwide, and early diagnosis is critical to improving patient outcomes. Despite advances in diagnostic imaging, such as MRI and CT, challenges still need to be solved in terms of cost, long processing times, and the need for specialized expertise. Deep learning (DL) techniques, particularly convolutional neural networks (CNNs) and vision transformers (ViTs) have shown great promise in medical image analysis, particularly for segmentation and classification tasks. However, limitations still need to be improved across diverse clinical datasets, as well as the need for large annotated datasets to train models effectively.

The primary objective of this thesis is to develop a robust computer-aided diagnosis (CAD) system that improves the detection and diagnosis of CVDs by utilizing advanced DL techniques, such as ensemble learning, transfer learning (TL), and ViTs. To address these challenges, this thesis introduces two key contributions: FCTransNet, an ensemble framework that combines multiple ViT models to improve segmentation accuracy and 2-TLViT, an enhanced transfer learning approach to optimize ViT models for CVD diagnosis. FCTransNet uses an Intelligent Weighted Summation Technique (IWST) to combine the outputs of individual ViT models to improve segmentation performance in cine MRI. The 2-TLViT approach incorporates both network-based and instance-based TL to improve generalization and reduce the need for large annotated datasets.

The methods used in this research involve the application of these techniques to cardiovascular imaging datasets, focusing on the segmentation of cardiovascular images. Extensive evaluations of FCTransNet show that it outperforms current ViT-based methods and sets a new benchmark for cardiac segmentation. The 2-TLViT approach demonstrates significant improvements in segmentation accuracy and model efficiency, with the use of a tailored weighted loss function helping to address the class imbalance in medical image segmentation tasks.

The results of this thesis highlight the potential of advanced DL techniques in enhancing CAD systems for CVD diagnosis. The proposed methods provide significant improvements in segmentation accuracy, generalization, and training efficiency, suggesting that they can help address current limitations in CAD systems and improve their clinical applicability. This research provides practical solutions for improving CVD diagnosis in various clinical settings by enhancing the accuracy, robustness, and generalization of CAD systems.

Keywords: Cardiovascular Diseases, Deep Learning, Vision Transformers, Cardiovascular Image Segmentation, Ensemble Learning, Transfer Learning.

الملخص

تظل الأمراض القلبية الوعائية واحدة من الأسباب الرئيسية للوفاة في العالم، ويعد التشخيص المبكر أمرًا بالغ الأهمية لتحسين نتائج المرضى. على الرغم من التقدم في تقنيات التصوير التشخيصي، مثل التصوير بالرنين المغناطيسي والأشعة المقطعية، لا تزال هناك تحديات يجب معالجتها فيما يتعلق بالتكلفة، وفترات المعالجة الطويلة، والحاجة إلى الخبرة المتخصصة. أظهرت تقنيات التعلم العميق(Deep Learning - DL) ، لا سيما الشبكات العصبية الالتفافية (Convolutional Neural Networks – CNN) والمحوّلات البصرية(Vision Transformers - ViTs) ، وعدًا كبيرًا في تحليل الصور الطبية، وخاصة في مهام التصنيف والتجزئة .ومع ذلك، لا تزال هناك محدوديات يجب تحسينها عبر مجموعات بيانات سريرية متنوعة، بالإضافة إلى الحاجة إلى مجموعات بيانات موسعة ومعلّمة لتدريب النماذج بشكل فعال. الهدف الرئيسي من هذه الرسالة هو تطوير نظام تشخيص مساعد يعتمد على الكمبيوتر لتحسين اكتشاف وتشخيص الأمراض القلبية الوعائية باستخدام تقنيات متقدمة من التعلم العميق، مثل التعلم بالأنظمة المجمعة (Ensemble Learning)، والتعلم بالتنقل(Transfer Learning - TL) ، والمحوّلات البصرية .لمواجهة هذه التحديات، تقدم هذه الرسالة مساهمتين رئيسيتين: الأولى هيFCTransNet ، وهو إطار عمل مجمع يدمج عدة نماذج من المحوّلات البصرية لتحسين دقة التجزئة، والثانية هي TLViT2-، وهي نهج محسّن للتعلم بالتنقل لتحسين نماذج المحوّلات البصرية في تشخيص الأمراض القلبية الوعائية. يستخدم FCTransNet تقنية متقدمة للدمج الذكي للصور تُسمى "تقنية الجمع الوزني الذكي (Intelligent Weighted Summation Technique - IWST) "لدمج نتائج نماذج المحوّلات البصرية الفردية وتحسين أداء التجزئة في التصوير بالرنين المغناطيسي السينمائي .(Cine MRI) يتضمن نهج TLViT 2-تقنيات من التعلم بالتنقل القائمة على الشبكات (Network-Based Transfer Learning) ومعتمدة على الحالات (Instance-Based Transfer Learning) لتحسين التعميم وتقليل الحاجة إلى مجموعات بيانات موسعة ومعلّمة. الأساليب المستخدمة في هذه البحث تتضمن تطبيق هذه التقنيات على مجموعات بيانات التصوير القلبي الوعائي، مع التركيز على تجزئة الصور القلبية الوعائية .أظهرت التقييمات الشاملة لـ FCTransNet أنه يتفوق على الأساليب الحالية القائمة على المحوّلات البصرية ويضع معيارًا جديدًا لتجزئة .كما يظهر نهج TLViT 2-تحسنًا كبيرًا في دقة التجزئة وكفاءة، حيث يساعد استخدام دالة فقدان مخصصة في معالجة التوازن بين الفئات في مهام تجزئة الصور الطبية. تسلط نتائج هذه الرسالة الضوء على إمكانيات تقنيات التعلم العميق المتقدمة في تعزيز أنظمة التشخيص المساعد باستخدام الكمبيوتر في تشخيص الأمراض القلبية الوعائية. توفر الطرق المقترحة تحسينات كبيرة في دقة التجزئة، والتعميم، وكفاءة التدريب، مما يشير إلى أنها يمكن أن تساعد في معالجة القيود الحالية في أنظمة التشخيص المساعد باستخدام الكمبيوتر وتحسين تطبيقاتها السريرية .تقدم هذه البحث حلولًا عملية لتحسين تشخيص الأمراض القلبية الوعائية في بيئات سريرية متنوعة، من خلال تعزيز الدقة، والمرونة، والتعميم في أنظمة التشخيص المساعد باستخدام الكمبيوتر.

الكلمات المفتاحية: الأمراض القلبية الوعائية، التعلم العميق، المحوّلات البصرية، تجزئة الصور القلبية الوعائية، التعلم بالأنظمة المجمعة، التعلم بالتنقل.

Résumé

Les maladies cardiovasculaires figurent parmi les principales causes de mortalité dans le monde, et un diagnostic précoce est essentiel pour améliorer les résultats des patients. Malgré les progrès des techniques d'imagerie diagnostique, telles que l'imagerie par résonance magnétique (MRI), des défis persistent en matière de coût, de durée de traitement et de nécessité d'expertise spécialisée. Les techniques d'apprentissage profond (DL), notamment les réseaux de neurones convolutionnels (CNNs) et les transformateurs de vision (ViTs), ont montré un grand potentiel dans l'analyse des images médicales, en particulier pour les tâches de segmentation. Cependant, des limitations subsistent, notamment la nécessité d'améliorer les modèles sur des ensembles de données cliniques diversifiés et le besoin de grands ensembles de données annotées pour un entraînement efficace des modèles.

L'objectif principal est de développer un système de diagnostic assisté par ordinateur (CAD) performant en utilisant des techniques avancées d'apprentissage profond, telles que l'apprentissage par transfert, les approches ensemblistes et les transformateurs de vision, pour améliorer la détection des maladies cardiovasculaires. Cette thèse présente deux contributions principales : FCTransNet, une approche ensembliste combinant plusieurs modèles ViTs pour améliorer la segmentation, et 2-TLViT, une méthode d'apprentissage par transfert pour optimiser les modèles ViTs. FCTransNet utilise une technique de fusion d'images (IWST) pour améliorer la segmentation en imagerie par résonance magnétique cinétique, tandis que 2-TLViT améliore la généralisation et réduit le besoin de grandes bases de données annotées.

Les évaluations approfondies de FCTransNet montrent qu'il surpasse les méthodes existantes basées sur les transformateurs de vision et établit une nouvelle référence pour la segmentation cardiaque. L'approche 2-TLViT montre des améliorations significatives en termes de précision de segmentation et d'efficacité des modèles, grâce à l'utilisation d'une fonction de perte pondérée sur mesure permettant de traiter le déséquilibre des classes dans les tâches de segmentation d'images médicales.

Les méthodes proposées permettent d'améliorer significativement la précision de la segmentation, la généralisation et l'efficacité de l'entraînement, suggérant qu'elles peuvent contribuer à surmonter les limitations actuelles des systèmes de diagnostic assisté par ordinateur et à améliorer leur applicabilité clinique. Cette recherche propose des solutions concrètes pour améliorer le diagnostic des maladies cardiovasculaires dans divers environnements cliniques, en renforçant la précision, la robustesse et la généralité des systèmes de diagnostic assisté par ordinateur.

Mots-clés: Maladies cardiovasculaires, apprentissage profond, transformateurs de vision, segmentation d'images cardiovasculaires, apprentissage par ensemble, apprentissage par transfert.

To my family.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, **Pr. Nabiha Azizi**, for her unwavering support, guidance, and availability throughout this journey. Her insightful advice and encouragement have been invaluable in shaping this work.

I extend my heartfelt thanks to my co-supervisor, **Dr. Samir Brahim Belhaouari**, for his assistance and constructive feedback. His expertise and dedication have significantly contributed to the progress and completion of this thesis.

I am sincerely grateful to the committee members for accepting the invitation to review my thesis. Their time and valuable feedback will greatly enrich this work.

Special thanks go to my friend, **Dr. Amel Slim**, for her guidance, the experiences she so generously shared, and her constant availability whenever I needed support. Her advice has been a beacon throughout my academic journey.

I owe an immense debt of gratitude to my family—my **mom, dad**, and siblings (**Foufa, Karrouma, Fayfou, Bolbol**, and the most adorable one, **Rahmou**)—for their unconditional love, encouragement, and support. They have always been my rock, and their belief in me has kept me going during the most challenging times.

Lastly, I would like to thank my dear friends from the student dormitory, **Rym Amata** and **Naouel Manaa**, for the unforgettable moments we shared during this journey. Their companionship and the beautiful memories we created together will always hold a special place in my heart.

To all those who contributed to this work in one way or another, I am profoundly thankful. This achievement would not have been possible without your support.

Contents

Abstract i

Résumé iii

Acknowledgements v

List of Figures ix

List of Tables xi

Abbreviations xii

1 Introduction 1

2 Cardiovascular Diseases and Imaging 6
2.1 Introduction . . . 6
2.2 Cardiovascular System Overview . . . 6
2.2.1 Cardiovascular System: The Heart . . . 6
2.2.1.1 Anatomy of the Heart . . . 7
2.2.1.2 The Cardiac Cycle . . . 9
2.2.2 Cardiovascular Diseases . . . 11
2.2.2.1 Definition . . . 11
2.2.2.2 Classification of Cardiovascular Diseases . . . 12
2.2.2.3 Epidemiology . . . 15
2.3 Imaging and Computer-Aided Diagnosis in CVDs . . . 19
2.3.1 Diagnosis of CVDs . . . 19
2.3.1.1 Clinical Evaluation . . . 19
2.3.1.2 Diagnostic Tests . . . 20
2.3.2 Cardiac Imaging Techniques . . . 20
2.3.2.1 Cardiac Magnetic Resonance Imaging . . . 21
2.3.3 Computer-Aided Diagnosis of CVDs . . . 23
2.3.3.1 Artificial Intelligence in Cardiovascular Diagnostics . . . 23
2.4 Conclusion . . . 32

3 Deep Learning Techniques for Medical Image Analysis 33

3.1 Introduction . . . 33
3.2 General Process of an Image-Based CAD System . . . 34
3.2.1 Preprocessing . . . 35
3.2.2 Assisted Diagnosis . . . 35
3.2.2.1 Classification . . . 36
3.2.2.2 Segmentation . . . 36
3.2.3 Decision . . . 41
3.3 Deep Learning Techniques for Medical Image Analysis . . . 42
3.3.1 Deep Learning Concepts . . . 42
3.3.2 Convolutional Neural Networks . . . 43
3.3.2.1 Fundamental Components of CNN . . . 44
3.3.2.2 Overview of CNN Architectures . . . 45
3.3.3 Vision Transformers . . . 48
3.3.3.1 Fundamental Components of ViT . . . 49
3.3.3.2 ViT Architectures in Image Segmentation . . . 51
3.3.4 Optimization of Deep Learning Models . . . 56
3.3.4.1 Transfer Learning . . . 57
3.4 Conclusion . . . 58

4 Intelligent Mask Image Reconstruction for Cardiac Image Segmentation through Local-Global Fusion 59
4.1 Introduction . . . 59
4.2 Related Works . . . 60
4.3 FCTransNet . . . 62
4.3.1 ROI Extraction Module . . . 63
4.3.2 Base Models of FCTransNet . . . 65
4.3.3 IWST-Based Fusion Module . . . 66
4.3.3.1 Intelligent Weighted Summation Technique . . . 66
4.4 Experimental Results . . . 68
4.4.1 Dataset . . . 68
4.4.2 Implementation Details . . . 69
4.4.3 Experimental Results and Ablation Studies . . . 70
4.4.3.1 Comparative Analysis with Related Works . . . 70
4.4.3.2 Analysis Study . . . 75
4.4.3.3 Ablation Experiments . . . 77
4.5 Discussion . . . 84
4.6 Conclusion . . . 87

5 Improved Two-stage Transfer Learning Approach for ViT-Based Myocardial Infarction Detection 88
5.1 Introduction . . . 88
5.2 Related Works . . . 89
5.3 Two-Stage Transfer Learning Framework . . . 91
5.3.1 Datasets and Preprocessing . . . 91
5.3.1.1 ACDC Datasets . . . 91
5.3.1.2 MyoPS Dataset . . . 93
5.3.2 Pretraining Phase . . . 94

5.3.2.1 Stage 1: Pretraining on the Classification Task 94
5.3.2.2 Stage 2: Pretraining on the Segmentation Task 95
5.3.3 Finetuning Phase 96
5.3.4 Weighted Loss Function 97
5.4 Experimental Results 98
5.4.1 Segmentation Results 98
5.4.1.1 Quantitative Evaluation 98
5.4.1.2 Qualitative Evaluation 99
5.4.1.3 Ablation Study 100
5.4.1.4 Analysis Experiments 104
5.4.1.5 Comparative Results 107
5.5 Discussion 107
5.6 Conclusion 109

6 Conclusions and Perspectives **110**

Bibliography **114**

List of Figures

2.1 Position of the heart in the thorax cavity. 7
2.2 Blood circulation through the heart. 8
2.3 The pericardial membranes and the layers of the heart wall. 9
2.4 Internal anatomical structures of the heart. 10
2.5 The pericardial membranes and the layers of the heart wall. 11
2.6 The atherosclerosis. 12
2.7 A coronary angiogram of atherosclerotic coronary arteries. 13
2.8 Prevalence of CVDs in the United States for adults aged 20 and older. . . 16
2.9 Age-standardized global prevalence rates of CVDs. 16
2.10 Deaths caused by CVDs in the United States. 18
2.11 Age-standardized global mortality rates of CVDs. 18
2.12 A CMR scan showing the short-axis plane of the heart. 22
2.13 Cardiac imaging planes and their corresponding standard views. 23

3.1 General process of an image-based CAD system. 34
3.2 Basic MLP with a single hidden layer. 43
3.3 Convolutional operation. 44
3.4 Max-pooling operation. 44
3.5 The standard VGGNet architecture. 45
3.6 The standard U-Net architecture. 47
3.7 Architecture of the ViT. 49
3.8 The structure of Self-Attention and Multi-Head Self-Attention. 51
3.9 Illustration of the Swin Transformer block. 53
3.10 Overview of the SwinUNet architecture. 54
3.11 Overview of the SegFormer architecture. 55
3.12 Overview of the TransUNet architecture. 56

4.1 Overview of the FCTransNet framework. 64
4.2 The enhanced UNet architecture for the extraction of the ROI. 65
4.3 Illustration of IWST technique application. 67
4.4 Process of Extracting 2D Slices from a 3D Cardiac MRI Volume. 69
4.5 Segmentation results visualization. 74
4.6 Bar chart comparing the DSC for transformer-based methods. 74
4.7 Bar chart comparing DSC for other DL-based methods. 75
4.8 Line chart comparing IWST with other image fusion techniques. 77
4.9 Segmentation results visualization for IWST with various image fusion techniques. 78
4.10 Segmentation results visualization without ROI extraction module. 81

4.11 Segmentation results with and without ROI extraction module. 82
4.12 Segmentation performance of FCTransNet during the ED and ES phases. 85

5.1 Overview of the proposed TL-based framework. 92
5.2 Example of 2D slices taken from the same CMR volume before and after ROI extraction. 93
5.3 An overview of the two-stage TL approach. 96
5.4 Bar chart comparison of different models. 99
5.5 Segmentation results visualization by different state-of-the-art segmentation models. 100
5.6 Segmentation results visualization of the ablation study. 102
5.7 Line chart illustrating the DSC scores from the ablation experiments. . . 102

List of Tables

2.1 Overview of DL-based CAD systems in CVD diagnosis. 25

4.1 Segmentation accuracy of FCTransNet compared to ViT-based methods. . 72
4.2 Segmentation accuracy of FCTransNet compared to other DL methods. . 73
4.3 Segmentation results comparison for IWST with various image fusion techniques (ED). 76
4.4 Segmentation results comparison for IWST with various image fusion techniques (ES). 76
4.5 Segmentation results with and without ROI extraction module (ED). . . . 79
4.6 Segmentation results with and without ROI extraction module (ES). . . . 80
4.7 Segmentation results comparison for FCTransNet with the base models (ED). 83
4.8 Segmentation results comparison for FCTransNet with the base models (ES). 83

5.1 Quantitative results comparison with state-of-the-art segmentation models. 103
5.2 Quantitative results comparison with state-of-the-art segmentation models. 103
5.3 Class distribution and corresponding class weights 105
5.4 Performance comparison between weighted and non-weighted loss functions. 106
5.5 Performance comparison between freezing and unfreezing weights. 106
5.6 Performance comparison between pretraining on ImageNet dataset and ACDC datasets. 106
5.7 Comparison of Edema and Scar DSC scores with previous studies on the MyoPS dataset. 107

Abbreviations

AHA	**A**merican **H**ealth **A**ssociation
ANN	**A**rtificial **N**eural **N**etwork
ASSD	**A**verage **S**ymmetric **S**urface **D**istance
CAD	**C**omputer-**A**ided **D**iagnosis
CMR	**C**ardiac **M**agnetic **R**esonance Imaging
CNN	**C**onvolutional **N**eural **N**etwork
CVDs	**C**ardiovascular **D**iseases
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etwork
DSC	**D**ice **S**imilarity **C**oefficient
ECG	**E**lectro**c**ardio**g**ram
ED	**E**nd **D**iastole
ES	**E**nd **S**ystole
FCN	**F**ully **C**onvolutional **N**etwork
HD	**H**ausdorff **D**istance
IoU	**I**ntersection **o**ver **U**nion
IWST	**I**ntelligent **W**eighted **S**mmation **T**echnique
LF	**L**oss **F**unction
LV	**L**eft **V**entricle
MI	**M**yocardial **I**nfarction
ML	**M**achine **L**earning
MLP	**M**ulty-**L**ayer **P**erceptron
MSA	**M**ulti-head **S**elf **A**ttention
Myo	**Myo**cardium
MRI	**M**agnetic **R**esonance **I**maging

NHANES	**N**ational **H**ealth and **N**utrition **E**xamination **S**urvey
NLP	**N**atural **L**anguage **P**rocessing
ROI	**R**egion **O**f **I**nterest
RV	**R**ight **V**entricle
TL	**T**ransfer **L**earning
ViT	**Vi**sion **T**ransformer
WHO	**W**orld **H**ealth **O**rganisation

Chapter 1

Introduction

Background

Cardiovascular diseases (CVDs) remain one of the leading causes of global morbidity and mortality, accounting for approximately 30% of all deaths worldwide, according to the World Health Organization (WHO) [1]. Among the most common and severe forms of CVDs are coronary artery disease, stroke, and myocardial infarction (MI). The rising prevalence of risk factors such as hypertension, diabetes, obesity, and sedentary lifestyles, compounded by the ageing global population, has further intensified the burden of CVDs across both developed and developing countries. Early detection and timely intervention are critical in reducing the morbidity and mortality associated with these conditions, but diagnosing CVDs in their early stages remains challenging due to the often subtle and nonspecific nature of symptoms, as well as the complexity of the disease.

MI, commonly referred to as a heart attack, occurs when blood flow to a part of the heart muscle is obstructed, typically due to a blockage in the coronary arteries. This leads to damage to the heart muscle and, if untreated, can result in long-term complications, including heart failure. Early and accurate diagnosis of MI is essential for improving patient outcomes. However, detecting MI and other CVDs presents significant difficulties due to the variability in symptoms and the need for specialized diagnostic tests. In clinical practice, cardiac imaging plays a central role in diagnosing CVDs. Techniques such as magnetic resonance imaging (MRI), computed tomography (CT) scans, and echocardiography are routinely used to assess the heart's structure and function, helping clinicians detect abnormalities such as coronary artery blockages, myocardial damage, and other structural defects [2, 3]. Despite their utility, these imaging modalities have inherent limitations, including high costs, long processing times, and the need for specialized equipment and expertise, all of which can lead to human error, particularly in time-sensitive situations.

In light of these challenges, artificial intelligence (AI), particularly deep learning (DL), has emerged as a promising solution to enhance the diagnosis and management of CVDs [4, 5]. DL models, such as Convolutional Neural Networks (CNNs), have demonstrated great success in medical image analysis, enabling automated detection, segmentation, and classification of abnormalities within medical images [6]. For instance, CNNs are capable of identifying key regions of interest (ROIs), such as infarcted tissue or coronary artery blockages, with high accuracy. These models can also automate tasks like image segmentation, which involves the precise delineation of anatomical structures or pathological regions, a crucial step for accurate diagnosis and treatment planning [7]. As a result, AI has the potential to improve the speed and accuracy of diagnoses, reduce the reliance on manual interpretation, and streamline clinical workflows.

An important area of focus in AI-based medical image analysis is segmentation. Image segmentation is the process of identifying and delineating regions of interest, such as infarcted tissue or coronary artery blockages, within medical images. Accurate segmentation is crucial for precise diagnosis and treatment planning, as it provides clinicians with detailed information about the location and extent of abnormalities. Traditional CNN-based models have shown considerable success in segmentation tasks, but there are still challenges related to capturing fine-grained details and handling variability in patient anatomy .

A promising new approach in image segmentation is the use of Vision Transformers (ViTs). Unlike traditional CNNs, which rely on convolutional layers to process images, ViTs employ self-attention mechanisms that allow the model to capture long-range dependencies and contextual information across the entire image [8, 9]. This makes ViTs particularly suited for tasks like medical image segmentation, where understanding the global context of an image is crucial for accurately identifying and delineating regions of interest. Recent studies have demonstrated that ViTs outperform traditional CNNs in various segmentation tasks, making them an exciting new tool in the field of medical image analysis [10].

Problem Statement

Despite their promising capabilities, AI-based models face several challenges when applied to medical image analysis. One of the main limitations is the issue of model generalization. A deep learning model that excels at solving one specific problem may perform poorly on other tasks or in different contexts. For example, a model trained to detect myocardial infarction in a specific dataset may struggle when applied to a different dataset from another hospital or imaging modality. This is especially true when datasets are extended or when there are discrepancies in imaging quality, patient

demographics, or other variables. Moreover, even though DL models can achieve high accuracy in medical applications, they often require large annotated datasets and extensive computational resources for training, which can be a significant bottleneck in their development and deployment in clinical settings.

To address these challenges, researchers have begun exploring advanced techniques such as ensemble methods and transfer learning (TL). Ensemble methods combine the outputs of multiple models to improve overall performance, as each model may excel in different aspects of the problem, thus enhancing the robustness and accuracy of predictions. On the other hand, TL allows models to leverage prior knowledge gained from other tasks or large datasets, reducing the need for extensive new data and improving generalization across different datasets. Both approaches aim to overcome the limitations of single-model deep learning, enabling models to perform more reliably across diverse clinical scenarios.

Continuous advances in medical technology have led to significant advances in the diagnosis and treatment of CVD. However, current CAD systems for CVD diagnosis still face significant challenges that limit their full potential . Despite their promising applications in identifying abnormalities and assisting clinicians, these systems often struggle with issues such as dataset diversity, model generalization, long training times, and the need for large, high-quality datasets. As a result, CAD systems do not always perform reliably across different patient populations or imaging modalities, and they still require significant human expertise for accurate interpretation. The question that arises is: ***"How can advanced DL techniques, including ensemble learning, TL, and ViTs, be effectively integrated into medical image analysis to improve the accuracy and generalization of cardiovascular image segmentation for the diagnosis of CVDs in diverse clinical datasets?"***

Purpose of the Study and Contributions

The primary goal of this thesis is the development of a robust CAD system for the detection and diagnosis of CVDs. The thesis aims to address key challenges faced by current CAD systems by leveraging advanced techniques, such as ensemble learning and transfer learning, to enhance the performance, accuracy, and generalization capabilities of the system.

Through the integration of ensemble learning, which combines multiple models to capitalize on their strengths, and the adaptation of transfer learning techniques, which allow the system to improve performance with limited annotated data, this research focuses on making CAD systems more efficient, robust, and adaptable to a variety of clinical

settings. The ultimate objective is to create a CAD system that not only provides accurate diagnoses but also adapts effectively to diverse patient populations and imaging conditions, ensuring reliable and timely decision-making in clinical practice.
The main contributions of this thesis are:

- This work introduces FCTransNet, an innovative ensemble framework designed for the accurate segmentation of cardiac structures in cine MRI. FCTransNet integrates the strengths of three state-of-the-art ViT models, effectively combining their capabilities to improve segmentation performance. A key component of this framework is the Intelligent Weighted Summation Technique (IWST), an advanced pixel-level image fusion approach that merges the outputs from each ViT model to construct a final segmentation mask. By leveraging the complementary strengths of the individual models, IWST enhances the accuracy and robustness of segmentation. Extensive evaluations on the ACDC dataset demonstrate that FCTransNet outperforms existing ViT-based methods and other deep learning approaches, setting a new benchmark for segmentation in cardiac imaging.
- The second contribution of this thesis is the proposal of an improved two-stage TL approach designed to optimize ViT model layers and enhance segmentation performance while addressing the challenges of limited data and long training times. This approach combines network-based and instance-based transfer learning techniques, enabling the model to leverage knowledge from a classification dataset, thus improving segmentation accuracy and reducing training time. Additionally, an enhanced ViT model featuring a tailored weighted loss function is explored, specifically designed to address class imbalance in segmentation tasks, which is common in medical imaging. The proposed framework is applied to the diagnosis of myocardial infarction (MI) and rigorously evaluated to demonstrate its effectiveness in real-world clinical scenarios.

Outline of the Thesis

This manuscript is organized as follows: Chapter 2 provides a detailed introduction to CVDs, their prevalence, and their impact on global health. It provides a comprehensive overview of the cardiovascular system, covering cardiac anatomy and physiology, as well as common CVDs and their epidemiologic implications. It discusses various diagnostic methods for CVD, with a particular focus on imaging techniques such as MRI and CT. The chapter also examines the role of CAD systems in the detection and management of CVD, highlighting the potential of DL techniques to improve accuracy and efficiency. Section 2.2 focuses on the cardiovascular system and CVDs, while Section 2.3 examines

imaging modalities and CAD methods, particularly DL-based techniques for image classification and segmentation.

Chapter 3 examines the transformative impact of DL techniques on medical image analysis, particularly in the context of cardiac imaging. It begins by explaining the general process of an image-based CAD system, covering stages such as preprocessing, classification, segmentation, and decision-making. The chapter then moves to an in-depth exploration of key DL architectures, including CNNs and ViTs. It discusses their application to medical image segmentation tasks. The chapter also covers optimization strategies that are crucial for improving the performance of DL models, such as TL. Section 3.2 describes the general workflow of image-based CAD systems, while Section 3.3 provides an overview of DL techniques and optimization strategies.

Chapter 4 introduces the first main contribution of the thesis: FCTransNet, a novel ensemble framework for accurate cardiac segmentation in cine MRI. FCTransNet combines three state-of-the-art ViT models to leverage their strengths and improve the robustness and accuracy of segmentation. A key component of this framework is the Intelligent Weighted Summation Technique (IWST). This advanced image fusion approach intelligently combines the output of each ViT model to construct a final, highly accurate segmentation mask. Extensive evaluations of FCTransNet on the ACDC dataset demonstrate its superiority over existing ViT-based methods and other DL approaches, establishing a new benchmark for cardiac image segmentation.

Chapter 5 presents the second main contribution of the thesis: an enhanced two-stage transfer learning (2-TLViT) approach for optimizing ViT models in the context of CVD diagnosis. This approach combines network-based and instance-based transfer learning techniques that allow the ViT models to leverage knowledge from classification datasets, thereby improving segmentation performance and reducing reliance on large annotated datasets. In addition, a customized weighted loss function is explored to address class imbalance in segmentation tasks, which is a common challenge in medical imaging. The chapter demonstrates how this two-stage transfer learning approach can be applied to the diagnosis of MI and evaluates its effectiveness in real clinical scenarios, ultimately improving the efficiency and accuracy of the CAD system.

Finally, we conclude this thesis by discussing limitations and potential future works in Chapter 6.

Chapter 2

Cardiovascular Diseases and Imaging

2.1 Introduction

This chapter provides a comprehensive introduction to CVDs, which are central to the research contributions of this thesis. It includes critical details on the severity of these diseases, as well as anatomical and physiological information about the cardiovascular system. The chapter is organized into two main sections.
Section 2.2 provides a comprehensive overview of the cardiovascular system, including detailed descriptions of the heart's anatomy and physiology. It addresses prevalent and severe CVDs, presenting an epidemiological analysis to illustrate their impact on global morbidity rates. Additionally, it outlines various diagnostic methods for CVDs. Section 2.3 describes different imaging techniques and modalities used in cardiovascular disease diagnosis. It provides an overview of computer-aided diagnosis for these diseases, with a focus on deep learning (DL) techniques in cardiovascular image analysis, including both classification and segmentation tasks.

2.2 Cardiovascular System Overview

2.2.1 Cardiovascular System: The Heart

The cardiovascular system is composed of the heart, blood vessels, and blood. The heart is a muscular organ that functions as the central pump of the cardiovascular system. It is responsible for circulating blood throughout the body, supplying oxygen and nutrients

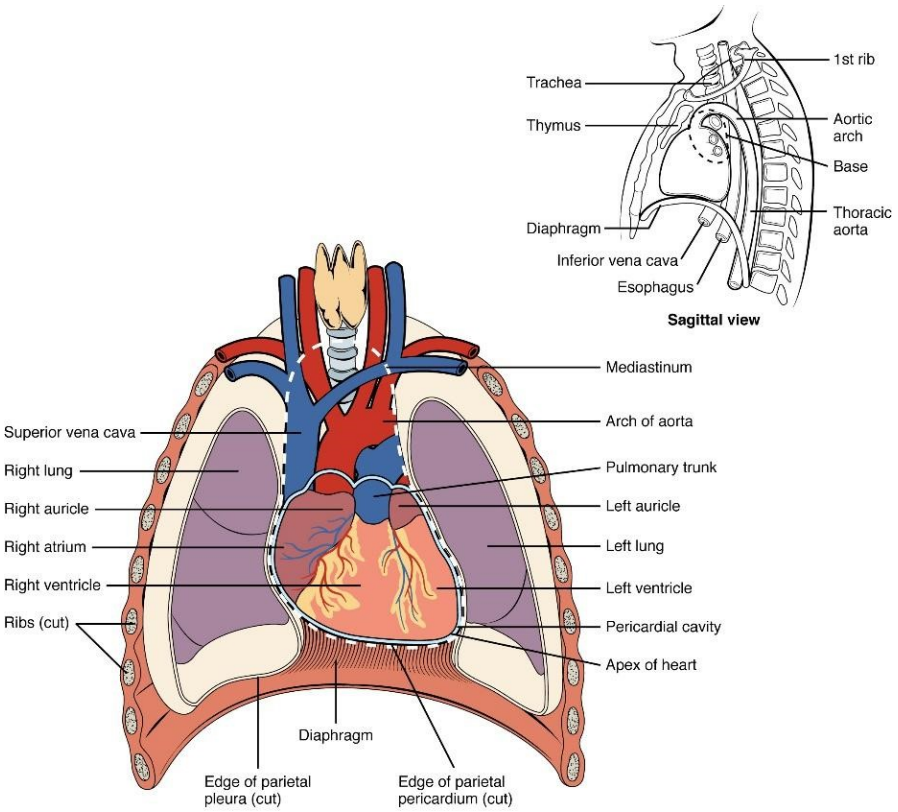

FIGURE 2.1: Position of the heart in the thorax cavity [11].

to tissues, and removing waste products. The heart is located in the thoracic cavity, specifically in the mediastinum, slightly left of the midline, and between the lungs. It rests on the diaphragm and is protected by the rib cage. The heart's position is crucial for its function, ensuring efficient blood flow to the entire body [11]. Figure 2.1 shows the location of the heart within the thoracic cavity.

2.2.1.1 Anatomy of the Heart

Circulation through the Heart

The heart is divided into four chambers: two upper chambers known as atria and two lower chambers known as ventricles. Blood enters the heart through the right atrium and gets pumped into the right ventricle (RV). From there, it travels to the lungs through the pulmonary arteries for oxygenation. The oxygen-rich blood returns to the heart via the left atrium and flows into the left ventricle (LV), which pumps it out to the body through the aorta. This organized blood flow ensures efficient oxygenation of tissues and the removal of metabolic waste products. Valves located between the atria and ventricles, as well as between the ventricles and the major arteries, prevent backflow and ensure

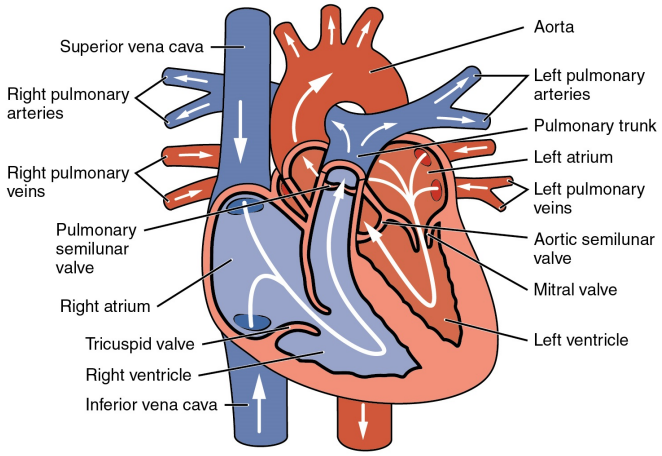

FIGURE 2.2: Blood circulation through the heart (Modified from [11]).

blood moves in one direction, thereby supporting effective circulation throughout the body. Figure 2.2 shows the blood circulation inside the heart.

Membranes and Layers

The heart is surrounded by a double-walled sac called the pericardium, which consists of two main layers: the fibrous pericardium and the serous pericardium [12]. The fibrous pericardium is the tough, inelastic outermost layer that anchors the heart to the surrounding structures and prevents it from overstretching. The serous pericardium is the inner layer and is further divided into the parietal layer, which lines the inner surface of the fibrous pericardium, and the visceral layer (also known as the epicardium), which adheres directly to the surface of the heart. Between the parietal and visceral layers is the pericardial cavity, which contains a small amount of pericardial fluid to reduce friction during heartbeats (Figure 2.3).
The heart wall itself is composed of three distinct layers [11]:

1. **Epicardium**: The outermost layer, also known as the visceral layer of the serous pericardium, which serves as a protective layer and often contains fat.

2. **Myocardium (Myo)**: The thick, muscular middle layer composed of cardiac muscle cells. This layer is responsible for the contractile force that pumps blood throughout the body.

3. **Endocardium**: The innermost layer, a smooth, thin membrane that lines the interior of the heart chambers and covers the heart valves. It provides a smooth surface for blood flow and helps prevent clot formation.

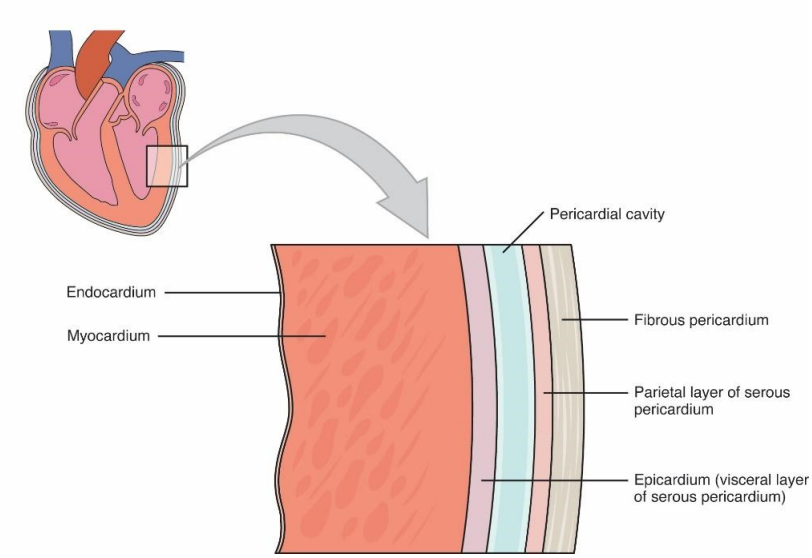

FIGURE 2.3: The pericardial membranes and the layers of the heart wall [11].

Internal Structure of the Heart

The four chambers of the heart are separated by muscular walls known as septa, which are crucial for maintaining the proper direction and separation of blood flow within the heart. The interatrial septum is a thin wall that separates the left and right atria, preventing the mixing of oxygenated blood from the left atrium with deoxygenated blood from the right atrium. During fetal circulation, there is an opening in this septum called the foramen ovale, which normally closes after birth. The interventricular septum is a thick, muscular wall that separates the left and right ventricles, preventing the mixing of oxygenated blood in the LV with deoxygenated blood in the RV. This septum is essential for maintaining efficient and effective separation of systemic and pulmonary circulations. The interventricular septum is divided into two parts: the muscular part, which is the lower, thicker portion composed primarily of muscle, and the membranous part, which is the upper, thinner portion more prone to congenital defects, such as ventricular septal defects. These septa are vital for the proper functioning of the heart, ensuring that oxygen-rich blood is efficiently separated from oxygen-poor blood, thus maintaining the efficacy of the circulatory system. Figure 2.4 illustrates the internal anatomical structure of the heart.

2.2.1.2 The Cardiac Cycle

The cardiac cycle consists of alternating phases known as *systole* and *diastole*. Systole refers to the contraction phase of the heart chambers (atria or ventricles), where blood is either pumped into the adjacent chamber or ejected into the arteries. At the end

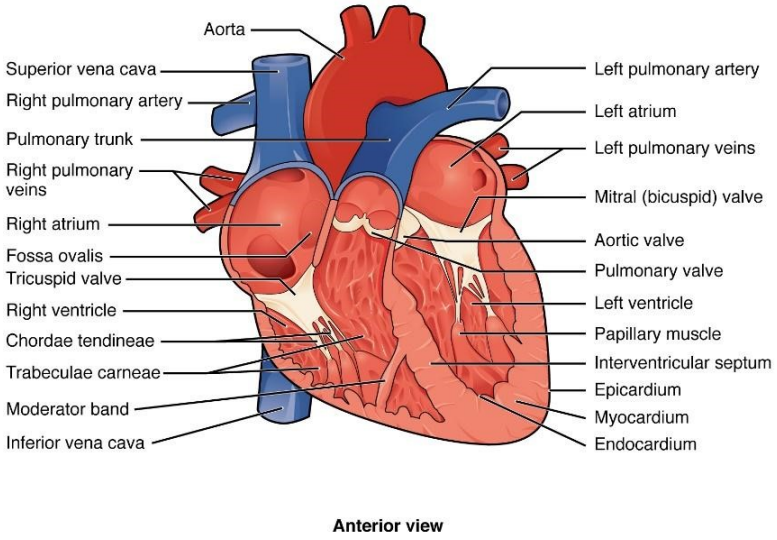

FIGURE 2.4: Internal anatomical structures of the heart [11].

of this contraction phase, the heart reaches the point known as end systole (ES), the moment when the volume of blood within the chamber is at its minimum. Diastole, on the other hand, is the relaxation phase where the heart chambers fill with blood. At the end of diastole, or end diastole (ED), the chamber volume is at its maximum before the onset of the next contraction. Both the atria and ventricles experience both contraction (systole) and relaxation (diastole), and precise coordination is crucial to ensure efficient blood pumping throughout the body [12]. The cardiac cycle is divided into four primary phases, each essential for effective blood circulation Figure 2.5:

- **Atrial Systole**: During this phase, the atrial muscles contract, increasing atrial pressure and pushing blood into the ventricles through open atrioventricular valves (tricuspid and mitral valves).
- **Atrial Diastole**: Following atrial systole, the atria relax (atrial diastole). Blood continues to flow passively from the veins into the atria, allowing for further filling of the ventricles. This phase accounts for the majority of ventricular filling.
- **Ventricular Systole**: Ventricular systole begins with the contraction of the ventricular muscles, causing the pressure within the ventricles to rise. Initially, this pressure is insufficient to open the semilunar valves, so no blood is ejected (isovolumic contraction). As the pressure increases further, it eventually surpasses the pressure in the pulmonary artery and aorta, causing the semilunar valves to open and blood to be pumped out of the ventricles (ventricular ejection). The LV generates higher pressure than the RV due to the greater resistance in the systemic circulation.

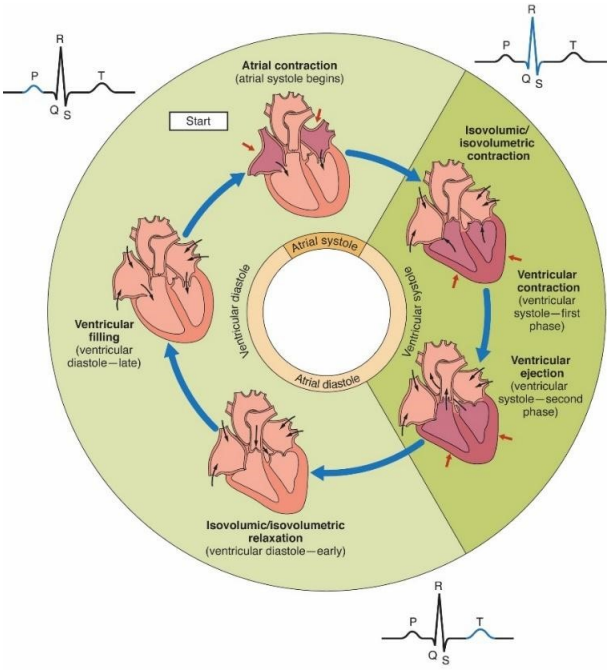

FIGURE 2.5: The pericardial membranes and the layers of the heart wall [11].

- **Ventricular Diastole**: Ventricular diastole follows with the relaxation of the ventricular muscles, leading to a decrease in ventricular pressure. In the early phase (isovolumic relaxation), the pressure drops below the pressure in the pulmonary artery and aorta, causing the semilunar valves to close and preventing backflow. As the pressure continues to fall, it eventually drops below the pressure in the atria, causing the atrioventricular valves to open. This allows blood to flow from the atria into the ventricles (late ventricular diastole), completing the cardiac cycle and preparing the heart for the next beat.

2.2.2 Cardiovascular Diseases

2.2.2.1 Definition

CVDs refer to a class of diseases that involve the heart or blood vessels, including conditions such as coronary artery disease, hypertension, heart failure, and arrhythmia. These diseases often result from atherosclerosis, characterized by the buildup of fatty deposits within the arterial walls. This leads to reduced blood flow and potential blockages (Figure 2.6). CVDs are the leading cause of morbidity and mortality globally, contributing to approximately 17.9 million deaths per year, according to the World Health Organization (WHO) [1].

Figure 2.6 illustrates the process of atherosclerosis and its impact on the coronary arteries. (a) depicts a normal artery with unobstructed blood flow and an artery with

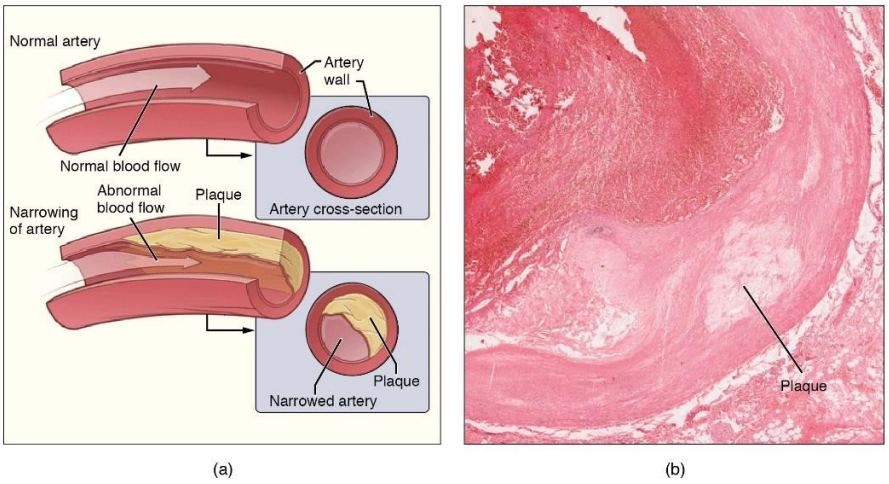

FIGURE 2.6: The atherosclerosis. (a) An illustration of the atherosclerosis. (b) A coronary artery showing the accumulation of connective tissue within the arterial wall [11].

plaque buildup, resulting in narrowing and abnormal blood flow. (b) provides a microscopic view of a coronary artery, highlighting the accumulation of connective tissue within the arterial wall, indicative of plaque buildup characteristic of atherosclerosis. This visual emphasizes how atherosclerosis reduces arterial lumen and impedes blood flow, contributing to CVDs.

2.2.2.2 Classification of Cardiovascular Diseases

Coronary Artery Disease

Coronary artery disease, also known as ischemic heart disease, is characterized by the narrowing or blockage of the coronary arteries due to the buildup of atherosclerotic plaques (Figure 2.7), which consist of cholesterol, fatty substances, cellular waste products, calcium, and fibrin [13]. This condition impairs blood flow to the heart muscle, leading to a deficiency in oxygen and nutrient supply. As a result, coronary artery disease can cause symptoms such as chest pain (angina), shortness of breath, and fatigue. Severe cases may result in a heart attack (myocardial infarction) when blood flow is completely obstructed.

Figure 2.7 is a coronary angiogram, which demonstrates the presence of significant atherosclerotic blockages in the coronary arteries. In particular, it illustrates the existence of blockages in the common trunk of the left coronary artery and the circumflex artery. These blockages result in a reduction of blood flow to the heart muscle, which can ultimately lead to severe consequences such as myocardial infarction and ischemic conditions. The left coronary artery supplies blood to a considerable proportion of the heart, and its obstruction can have a markedly detrimental impact on cardiac function.

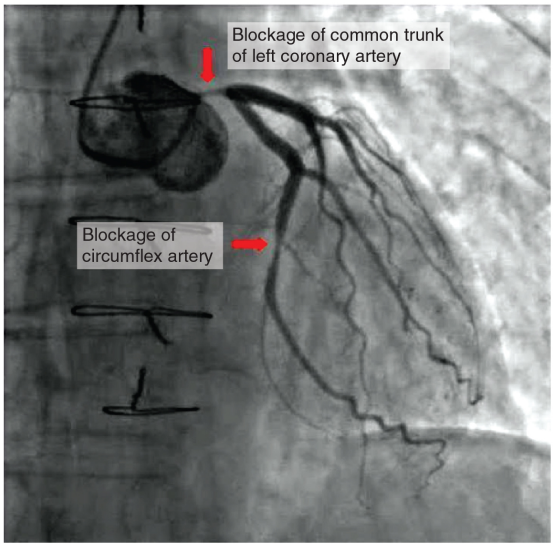

FIGURE 2.7: A coronary angiogram of atherosclerotic coronary arteries [11].

Myocardial Infarction

Myocardial infarction (MI), commonly known as a heart attack, is a critical cardiovascular event characterized by the sudden obstruction of blood flow to a segment of the heart muscle, typically due to the rupture of an atherosclerotic plaque and subsequent formation of a blood clot [14]. This occlusion results in the death of myocardial tissue, leading to the formation of scar tissue and the accumulation of edema within the affected area. Scar tissue replaces the necrotic muscle cells and impairs the heart's ability to contract effectively, contributing to decreased cardiac function and potential complications such as heart failure. Edema, or the accumulation of fluid due to inflammation, serves as an acute marker of tissue injury and is crucial for assessing the viability of myocardial tissue post-MI [15]. Advanced imaging techniques, such as cardiac magnetic resonance imaging (MRI), play a pivotal role in distinguishing between scarred, edematous, and healthy myocardial tissue, thereby guiding clinical management and treatment strategies aimed at minimizing further damage and promoting cardiac recovery [16]. Symptoms of MI include severe chest pain, shortness of breath, nausea, and lightheadedness. Myocardial infarction is a critical medical emergency that significantly contributes to cardiovascular morbidity and mortality worldwide.

Heart Failure

Heart failure is a clinical syndrome characterized by the heart's inability to pump blood efficiently to meet the body's metabolic demands. This condition typically arises from structural or functional abnormalities in the heart that impair its ability to contract or

relax effectively. Common causes include coronary artery disease, hypertension, valvular heart disease, and cardiomyopathies [17]. These conditions can lead to two main types of heart failure: Systolic Heart Failure occurs when the heart muscle becomes weakened and cannot contract forcefully enough during systole (the pumping phase) to eject sufficient blood into circulation. Diastolic Heart Failure is where the heart muscle becomes stiff and less compliant, impairing its ability to relax and fill adequately with blood during diastole (the filling phase). Heart failure is often progressive and can lead to symptoms such as fatigue, shortness of breath (especially with exertion or when lying down), and fluid retention. Diagnosis involves clinical evaluation, imaging techniques like echocardiography and MRI to assess heart structure and function, and sometimes invasive procedures like cardiac catheterization to measure pressures within the heart chambers.

Arrhythmia

The term 'arrhythmia' is used to describe a group of conditions characterized by the presence of abnormal heart rhythms. These may manifest as an increased heart rate (tachycardia), a decreased heart rate (bradycardia), or an irregular heart rhythm [18]. These irregularities can disrupt the heart's ability to pump blood effectively, which may manifest as symptoms such as palpitations, dizziness, shortness of breath, and chest pain or fainting. The etiology of arrhythmias is diverse, with underlying heart conditions such as coronary artery disease, electrolyte imbalances, structural abnormalities of the heart, and the adverse effects of certain medications all contributing to their development. The diagnosis of arrhythmias is typically achieved through electrocardiography, which enables the detection and classification of the arrhythmia in question. Additionally, other tests, such as Holter monitoring or electrophysiological studies, may be employed for further evaluation.

Hypertensive Heart Disease

Hypertensive heart disease encompasses a spectrum of conditions affecting the heart due to prolonged high blood pressure, known as hypertension. Chronic hypertension imposes increased strain on the heart muscle and blood vessels, leading to structural and functional changes over time. These changes include left ventricular hypertrophy, where the heart's main pumping chamber enlarges, and diastolic dysfunction, which impairs the heart muscle's relaxation. Ultimately, hypertension can progress to heart failure and increase the risk of coronary artery disease, resulting in conditions such as angina (chest pain) or myocardial infarction (heart attack). Hypertensive heart disease also involves

endothelial dysfunction and vascular remodeling, contributing to increased stiffness of the arteries and impaired coronary circulation. These factors further exacerbate the risk of myocardial ischemia and arrhythmias, highlighting the complex interplay between hypertension and cardiovascular pathology.

2.2.2.3 Epidemiology

Epidemiology is a fundamental tool in understanding the prevalence, distribution, and risk factors associated with CVDs, which is crucial for computational modeling and data-driven healthcare solutions. CVDs, encompassing conditions such as coronary artery disease, hypertension, heart failure, and stroke, present complex challenges influenced by genetic predispositions, environmental factors, and lifestyle choices [19]. This section explores epidemiological patterns in CVDs, including incidence rates across diverse populations and geographic regions, aiming to uncover actionable insights through computational analyses.

Prevalence and Incidence

CVDs represent a major global health challenge with substantial prevalence and incidence rates. In the United States, data from National Health and Nutrition Examination Survey (NHANES) [20] from 2017 to March 2020 indicate that the overall prevalence of CVD is 48.6% among adults aged 20 and older, which translates to approximately 127.9 million individuals (Figure 2.8). This prevalence increases with age and differs by sex. Excluding hypertension, the prevalence of coronary heart disease, heart failure, and stroke alone stands at 9.9%, affecting around 28.6 million people [21].

Globally, a study conducted on 56,716 adults aged 40 and older in northern China found that 22.7% had a high 10-year risk of CVD according to WHO/International Society of Hypertension risk prediction charts. This study also reported high age-adjusted prevalences of hypertension (54.3%), dyslipidemia (36.5%), obesity (24.8%), and diabetes (18.2%) [21].

Incidence rates further emphasize the burden of CVDs [22]. A meta-analysis of 32 studies involving Asian adults aged 18 to 92 years, who were free of CVD at baseline and followed for over 10 years, revealed an incidence of fatal CVD at 3.68 events per 1,000 person-years. Significant risk factors for long-term fatal CVD included male sex (risk ratio of 1.49), older age (risk ratio of 7.55), and current smoking (risk ratio of 1.68) [21]. This highlights the ongoing and substantial burden of CVDs [22], emphasizing the need for effective prevention and management strategies.

Figure 2.9 illustrates the age-standardized global prevalence rates of CVDs per 100,000

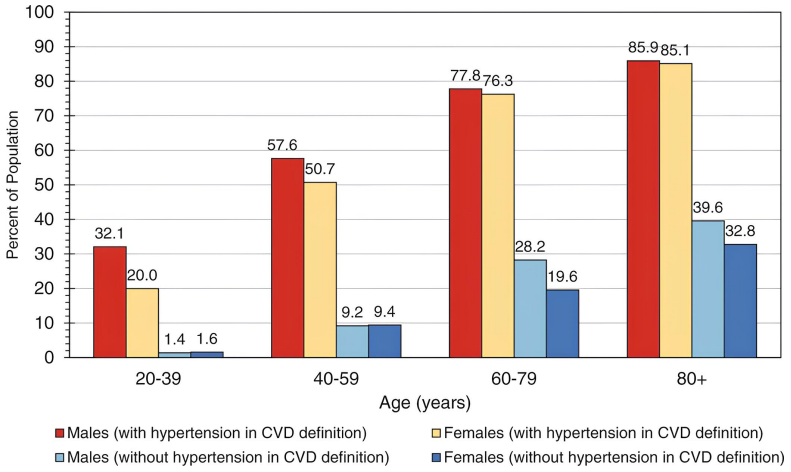

FIGURE 2.8: Prevalence of CVDs in the United States for adults aged 20 and older, by age and sex (Source: Unpublished National Heart, Lung, and Blood Institute tabulation using NHANES, 2017-2020 [20]).

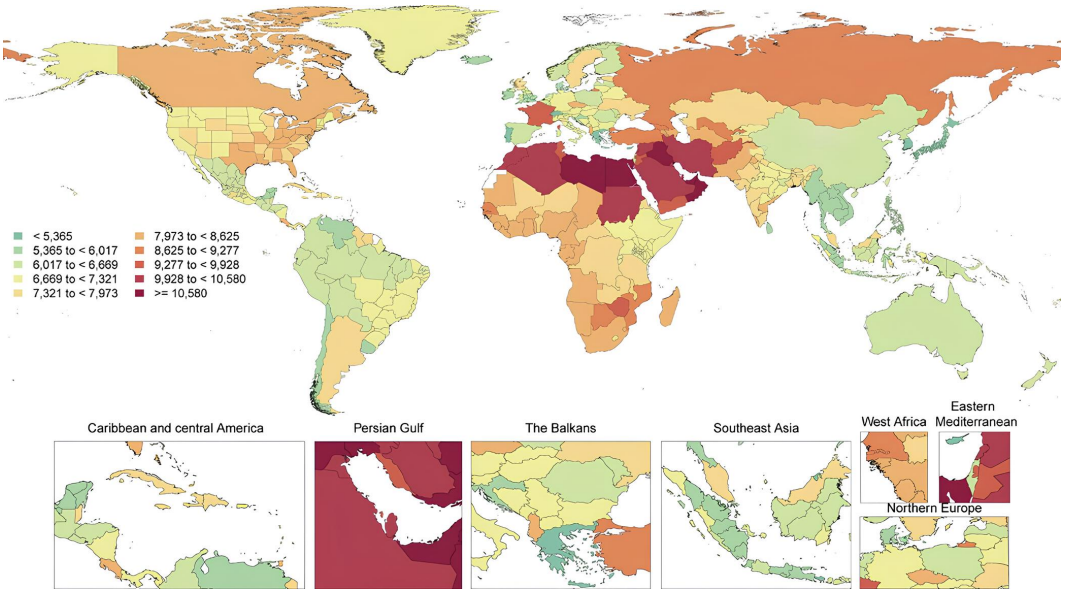

FIGURE 2.9: Age-standardized global prevalence rates of CVDs per 100,000 individuals, encompassing both sexes (Source: the American Health Association (AHA), 2021 [21]).

individuals for both sexes, using a color-coded scale. The highest rates are observed in the Persian Gulf and Eastern Mediterranean regions, indicating a considerable disease burden, potentially attributable to lifestyle factors and healthcare access. Moderate to high rates are observed in Southeast Asia, the Balkans, and parts of the Caribbean and Central America. Moderate prevalence is observed in Europe, Australia, and parts of South America, while the lowest rates are found in Western and Central Africa and Central Asia.

Mortality and Morbidity

CVDs remain a leading cause of global mortality and morbidity, with a complex and evolving impact over time [1], [21]. In the United States, deaths attributed to diseases of the heart and other cardiovascular conditions have shown a varied trajectory. Historically, these deaths increased steadily from the early 1900s through the 1980s, saw a decline into the 2010s, but have unfortunately risen again in recent years (Figure 2.10). As of 2021, coronary heart disease was the leading cause of CVD-related death, accounting for 40.3% of CVD fatalities, followed by stroke at 17.5%, other minor CVD causes combined at 17.1%, high blood pressure at 13.4%, heart failure at 9.1%, and diseases of the arteries at 2.6% [21].

The age-adjusted death rate for CVD increased from 228.6 per 100,000 people in 2011 to 233.3 per 100,000 in 2021, reflecting a 2.1% rise over this period. Notably, the Million Hearts 2022 Initiative aims to prevent 1 million heart attacks, strokes, and other cardiovascular events, underscoring the ongoing challenge of CVDs [23]. In 2016, there were over 1,000 deaths daily from heart attacks, strokes, or other cardiovascular events, with 2.2 million hospitalizations and 415,480 deaths related to CVD reported that year. Additionally, 35% of life-changing cardiovascular events occurred in adults aged 35 to 64 years, contributing to 775,000 hospitalizations and 73,000 deaths within this age group [21].

Geographic variation in CVD events is significant, with the Southeast and Midwest regions of the United States reporting the highest rates. Conversely, states like Utah, Wyoming, and Vermont have some of the lowest CVD event rates [21]. This geographic disparity illustrates the influence of regional factors on cardiovascular health outcomes (Figure 2.11).

Globally, CVDs claim more lives each year than cancer and chronic lower respiratory diseases combined [21]. In 2021, the global mortality rate for heart disease and stroke was 214.9 per 100,000 people. The total number of resident deaths in the United States. for that year reached 3,464,231, with CVD accounting for a substantial portion of these deaths. The rates of CVD deaths have fluctuated over the decades, with a notable increase in recent years, and disparities in mortality rates continue to exist among different racial and ethnic groups [21].

Figure 2.11 depicts the global mortality rates of CVDs per 100,000 individuals for both sexes, using a color-coded scale. The highest mortality rates, indicated by dark red, are predominantly observed in Eastern Europe, Central Asia, and parts of the Eastern Mediterranean region. Moderate to high mortality rates, shown in shades of orange and red, are observed in the Balkans, Southeast Asia, and the Persian Gulf. In contrast, regions such as North America, Western Europe, and Australia show relatively lower

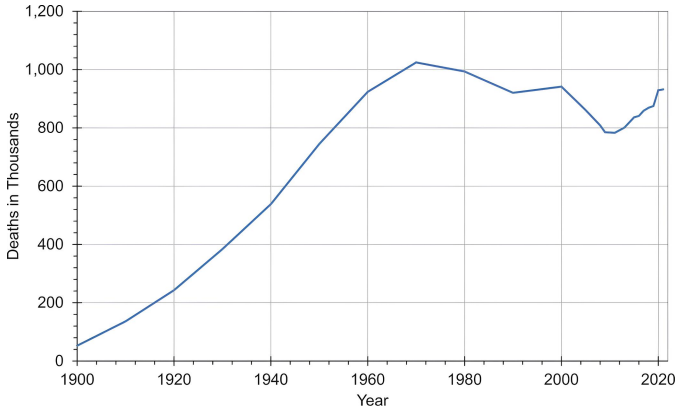

FIGURE 2.10: Deaths caused by CVDs in the United States (Source: Unpublished National Heart, Lung, and Blood Institute tabulation using National Vital Statistics System (NVSS), 1900-2021 [24]).

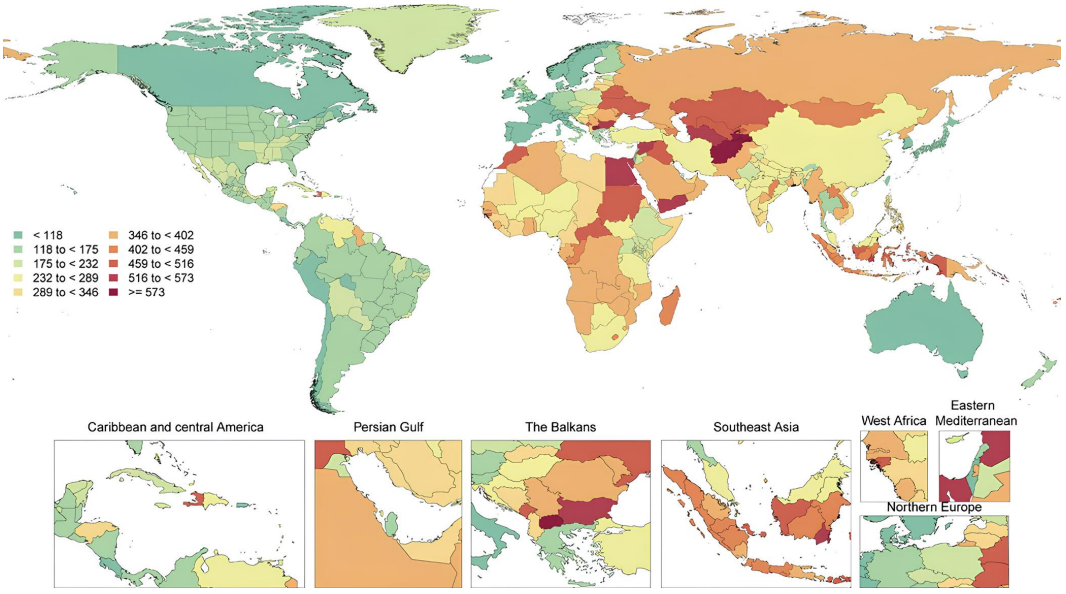

FIGURE 2.11: Age-standardized global mortality rates of CVDs per 100,000 individuals, encompassing both sexes (AHA, 2021 [21]).

mortality rates, marked in green and light yellow, suggesting the presence of more effective healthcare systems and preventive measures. Northern Europe and parts of Central America exhibit moderate mortality rates.

Demographic and Geographic Variations

The impact of CVDs varies significantly based on demographic factors and geographic location. Socioeconomic status, geographic region, and ethnicity all contribute to differences in CVD prevalence and outcomes. In lower-income countries, the prevalence of CVDs has been rising due to increased urbanization, lifestyle changes, and limited access to healthcare. For example, the highest CVD mortality rates are observed in Eastern

Europe and Central Asia, where socioeconomic factors and healthcare disparities contribute to higher rates of CVD (Figure 2.9).
In contrast, high-income countries, despite having better healthcare systems, still face significant CVD burdens due to lifestyle factors such as obesity and physical inactivity [22]. In regions like North America and Western Europe, there are also notable disparities in CVD outcomes based on socioeconomic status and access to healthcare services. Ethnic and racial differences also play a role in CVD prevalence. For instance, African American populations in the United States experience higher rates of hypertension and heart disease compared to other racial groups. Similarly, in parts of Asia, genetic predispositions and dietary factors contribute to varying rates of CVD.

2.3 Imaging and Computer-Aided Diagnosis in CVDs

2.3.1 Diagnosis of CVDs

The diagnosis of CVDs is a multifaceted process that involves a combination of clinical evaluation and diagnostic tests [25]. Accurate diagnosis is crucial for effective treatment and management, typically beginning with a thorough clinical evaluation followed by various diagnostic tests.

2.3.1.1 Clinical Evaluation

Clinical evaluation is the initial step in diagnosing CVDs and encompasses several critical components. A detailed patient history is fundamental in this process. Clinicians document symptoms such as chest pain, dyspnea, and palpitations, along with the patient's medical history, family history of CVDs, lifestyle factors including smoking, diet, and physical activity, and any previous cardiovascular events or treatments [17]. This comprehensive history is vital in revealing important risk factors and potential causes of cardiovascular issues.
Following the patient's history, a thorough physical examination is performed. Vital signs, including blood pressure, heart rate, respiratory rate, and temperature, are measured to detect abnormalities. The heart examination involves inspection, palpation, percussion, and auscultation to identify issues, such as murmurs, irregular rhythms, or abnormal heart sounds [26]. Additionally, a vascular examination is conducted to check peripheral pulses, examine jugular venous pressure, and assess for signs of peripheral artery disease [26]. The general examination also looks for physical signs that indicate cardiovascular conditions, such as edema, cyanosis, and xanthelasma.

2.3.1.2 Diagnostic Tests

After the clinical evaluation, various diagnostic tests are employed to confirm the diagnosis and assess the extent of CVDs. These tests include a range of non-invasive and invasive procedures designed to provide detailed information about the heart's condition.

Electrocardiograms (ECG) are among the most commonly used diagnostic tools. A resting ECG measures the electrical activity of the heart to detect arrhythmias, myocardial infarction, and other cardiac conditions [27]. For more continuous monitoring, Holter monitoring involves a continuous ECG recording over 24-48 hours to detect intermittent arrhythmias [28]. Stress testing, which involves an ECG performed during physical exertion, assesses the heart's response to stress and helps detect ischemia.

Imaging tests play a significant role in the diagnosis of CVDs. Echocardiography uses ultrasound to create images of the heart, allowing evaluation of heart structure, function, and blood flow [11]. Chest X-rays provide images of the heart, lungs, and chest cavity, helping to identify heart enlargement, pulmonary congestion, and other abnormalities. Advanced imaging techniques such as cardiac MRI and computed tomography (CT) scans offer detailed images of heart anatomy, function, and coronary arteries [2], [3].

Blood tests are also critical in diagnosing CVDs. Biomarkers such as cardiac troponins, BNP (B-type natriuretic peptide), and CRP (C-reactive protein) are measured to detect myocardial injury, heart failure, and inflammation [29]. A lipid profile is assessed to evaluate cholesterol and triglyceride levels, providing insight into the risk of atherosclerosis and coronary artery disease [30].

Non-invasive tests such as the ankle-brachial index and pulse oximetry are used to diagnose peripheral artery disease and measure oxygen saturation in the blood, respectively [31]. Invasive tests, including cardiac catheterization and coronary angiography, involve threading a catheter through blood vessels to the heart to measure pressures, take blood samples, and perform angiography to visualize coronary arteries [32].

Electrophysiological studies, which evaluate the electrical system of the heart, are used to diagnose and treat arrhythmias [33]. Each diagnostic test provides specific information about the heart's condition, and together, they form a comprehensive picture of cardiovascular health, aiding in the accurate diagnosis and management of CVDs.

2.3.2 Cardiac Imaging Techniques

Cardiac imaging techniques are fundamental tools in the diagnosis, assessment, and management of CVDs. These modalities provide detailed insights into the structure and function of the heart, which facilitate the identification of pathologies, the formulation of

treatment plans, and the monitoring of therapeutic progress. Various imaging techniques are employed in clinical practice, including echocardiography, CT, nuclear cardiology, and angiography. Each of these techniques offers unique insights and diagnostic capabilities. Echocardiography employs ultrasound technology to generate real-time images of the heart, enabling the assessment of chamber dimensions, wall motion, and valve functionality [34]. CT, particularly coronary CT angiography, provides detailed cross-sectional images of the coronary arteries and cardiac structures [3]. Nuclear cardiology utilizes radioactive tracers for the evaluation of myocardial perfusion and viability. Concurrently, angiography remains the gold standard for visualizing the coronary arteries and guiding interventions (Figure 2.7). Among these, cardiac MRI is particularly noteworthy for its superior soft tissue contrast and comprehensive assessment of cardiac morphology and function, rendering it indispensable in both clinical and research settings [2].

2.3.2.1 Cardiac Magnetic Resonance Imaging

Cardiac Magnetic Resonance Imaging (CMR) or MRI is a non-invasive imaging modality that uses powerful magnetic fields and radiofrequency waves to produce detailed images of the heart and its structures. It is highly valued for its exceptional soft tissue contrast and ability to provide comprehensive assessments of cardiac morphology, function, and tissue characterization [2]. CMR allows for precise visualization of the heart's chambers, myocardium, and blood vessels, making it a critical tool in diagnosing and monitoring various CVDs. Its versatility includes evaluating myocardial viability, detecting ischemic heart disease, and assessing cardiomyopathies. CMR's ability to acquire images in multiple planes without ionizing radiation underscores its significance in modern cardiac imaging.
Figure 2.12 demonstrates the efficacy of CMR imaging in providing a clear visualization of cardiac structures. The short-axis plane of the heart is displayed during the end-diastolic and the end-systolic phases. In the image (i), the heart at end-diastole displays the LV (blue), RV (red), and Myo (green) in their most dilated state, with each structure distinctly visible. Image (ii) depicts the heart at end-systole, focusing on the ventricles and myocardium following contraction. The CMR scan effectively demonstrates the changes in ventricular volumes and myocardial thickness between the two phases, showcasing its capability to provide detailed and precise anatomical and functional information about the heart.

CMR images are typically acquired in several standard planes to provide a comprehensive view of the heart's structure and function. These planes include:

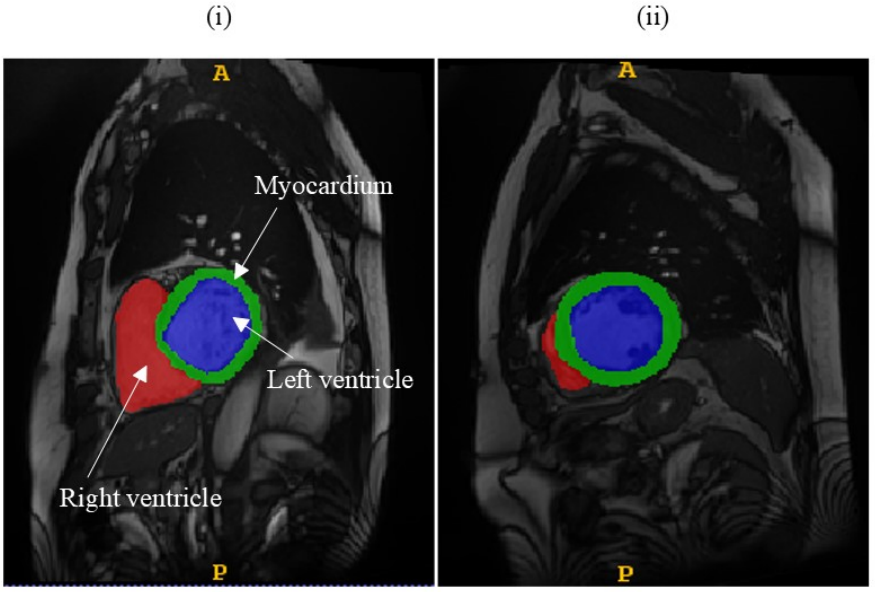

FIGURE 2.12: A CMR scan showing the short-axis plane of the heart. (i) end-diastolic phase and (ii) end-systolic phase.

- **Short-axis plane:** This plane slices the heart horizontally, providing cross-sectional images from the base to the apex of the heart. It is particularly useful for assessing the function and volume of the ventricles and for detecting myocardial infarctions.
- **Horizontal long-axis plane:** Also known as the four-chamber view, this plane slices the heart horizontally, capturing both atria and ventricles. It is essential for assessing overall cardiac function, chamber sizes, and detecting congenital heart defects.
- **Vertical long-axis plane:** This plane, sometimes known as the two-chamber view, slices the heart vertically, including the LV and left atrium. It is valuable for evaluating the left ventricular function and mitral valve abnormalities.

Figure 2.13 shows the standard cardiac imaging planes and the corresponding views. In panel (a), three standard imaging planes are illustrated, which align with the major axes of the heart: the horizontal long-axis, the vertical long-axis, and the short-axis. These planes intersect with important structures, including the aorta, pulmonary trunk, LV, and RV. Panel (b) depicts images captured along these planes. The horizontal long-axis (4-chamber view) shows all four heart chambers. The vertical long-axis (2-chamber view) displays a selected ventricle and its associated atrium. The short-axis view offers a frontal perspective of the Myo, highlighting the left and right ventricles (Figure 2.12).

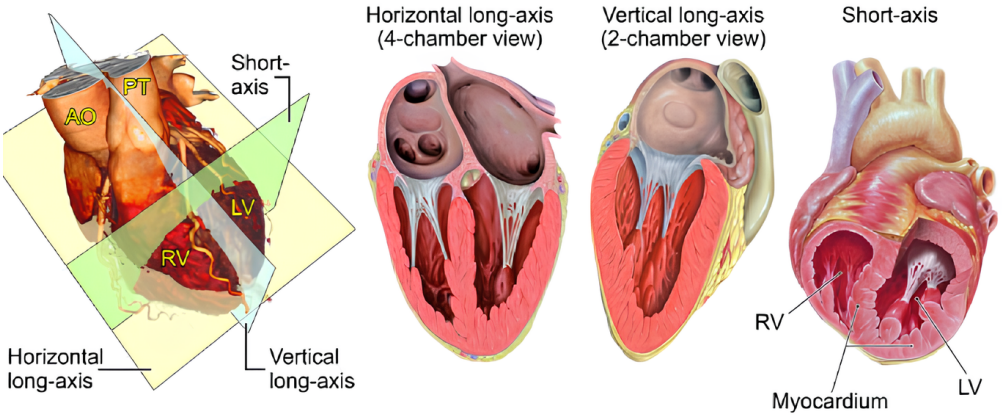

FIGURE 2.13: Cardiac imaging planes and their corresponding standard views. (a) Illustration of the three primary cardiac imaging planes: horizontal long-axis, vertical long-axis, and short-axis. (b) Images taken along these planes provide views of all four heart chambers [35].

2.3.3 Computer-Aided Diagnosis of CVDs

Traditional diagnostic methods, while still effective, present certain limitations [6]. Challenges include the potential for human error, variability in interpretation, and the growing complexity of medical data. To address these issues, computer-aided diagnosis (CAD) systems have emerged as a powerful tool in cardiovascular medicine [36].
CAD systems utilize advanced computational techniques, including artificial intelligence (AI) [37], [4], and image processing [38], to support clinicians in diagnosing CVDs with enhanced precision and efficiency [5]. These systems process vast amounts of medical data, encompassing imaging studies [39], [7], electrocardiograms [40], [41], and patient records, to detect patterns and abnormalities that may signify the presence of disease.
Integrating CAD systems into cardiovascular diagnostics provides numerous benefits [42]. First, they offer a standardized approach to interpretation, reducing inter-observer variability and enhancing diagnostic consistency. Additionally, CAD systems can analyze data rapidly, enabling faster decision-making in clinical settings [6]. Furthermore, these systems improve the sensitivity and specificity of diagnostic methods, which can lead to earlier disease detection and potentially better patient outcomes [43].

2.3.3.1 Artificial Intelligence in Cardiovascular Diagnostics

AI has emerged as a transformative tool in cardiology, offering significant advancements in diagnostic precision, risk assessment, and personalized treatment planning [44]. AI technologies, particularly machine learning (ML) and DL have greatly contributed to cardiovascular decision-making by enabling sophisticated analysis of complex clinical

data.

ML algorithms, including traditional techniques like logistic regression, support vector machines (SVMs), and random forests, are pivotal for analyzing cardiovascular data, identifying patterns, and predicting outcomes. These methods support early detection of CVDs, such as myocardial infarction and other cardiac events, by leveraging patient-specific data [45]. While traditional ML approaches provide robust diagnostic and risk assessment tools, they often face limitations when dealing with high-dimensional data and capturing complex patterns in cardiovascular datasets [46]. Scalability and feature extraction issues further constrain traditional ML models, making it challenging to address the intricacies of medical data comprehensively.

In contrast, DL techniques offer enhanced capabilities for handling high-dimensional and intricate cardiovascular data due to their ability to automatically learn and extract relevant features from large, diverse datasets [47]. Leveraging multi-layer neural networks, DL has transformed the analysis of medical images, ECGs, and electronic health records [48]. Advanced architectures like convolutional neural networks (CNNs), generative adversarial networks (GANs), autoencoders (AEs), and Transformers have been especially impactful, supporting diverse cardiovascular tasks such as disease classification, segmentation, anomaly detection, image reconstruction, and automated report generation.

Table 2.1 provides a comprehensive summary of recent advances in DL-based approaches for CAD of CVDs, with a special focus on imaging data. Each study is thoroughly detailed, beginning with the name of the proposed approach or model, followed by its specific application in CVD diagnosis, such as classification, segmentation, anomaly detection, image reconstruction, and automated report generation. The table then lists the datasets used in each study, reflecting the robustness and generalizability of these approaches to different cardiovascular imaging data. Imaging modalities such as MRI, CT, and echocardiograms are identified, highlighting the range of diagnostic inputs analyzed by these models.

The table also outlines any preprocessing techniques applied to the data, highlighting efforts to improve model accuracy and reliability. The core DL architectures and training strategies used in each approach are described, highlighting advanced models such as CNNs, GANs, RNNs, AEs, and Transformers that have shown significant efficacy in handling high-dimensional and complex cardiovascular data. Finally, the table presents the performance metrics used in each study, detailing key evaluation metrics such as accuracy, sensitivity, specificity, and other evaluation criteria that validate the effectiveness of these CAD systems in supporting accurate and timely diagnosis of CVD. This structured review allows for a critical comparison of methodologies, datasets, and results, providing insight into the state-of-the-art in DL-based CAD for CVDs applications.

Table 2.1: Overview of DL-based CAD systems in CVD diagnosis.

Reference	Approach Name	Application	Dataset	Imaging Modality	Data Preprocessing	DL Technique and Training Strategy	Test Performance
Islam et al. [49]	CNN-MCD	ECG images classification	Cardiac and COVID-19 dataset [50]	ECG	Convert images to grayscale, resize to 70×70, crop	CNN combined with Monte Carlo Dropout, MCD, train 80% val 10% test 10%	**Accuracy:** 93.90% **Precision:** 94.00% **Sensitivity:** 93.00%
Oksuz et al. [51]	/	Classification for motion artifacts	UK Biobank [52]	MRI	Data augmentation using k-space based training approach	3D spatiotemporal CNN, LF: binary cross-entropy, 10-fold cross validation	**Accuracy:** 0.982 **Precision:** 0.809 **Recall:** 0.652 **F1 score:** 0.704
Zreik et al. [53]	/	Detection and characterization of coronary artery plaque type	Private	MRI	Data augmentation (random rotations)	Multi-task recurrent CNN, LF: categorical cross-entropy, train 50% val 10% test 40%	**Accuracy:** 0.77–0.80 **F1 score:** 0.61–0.75 **Cohen's kappa:** 0.61–0.68
Liu et al. [54]	EDMAE	Classification of pediatric cardiac ultrasound	Private	IVUS	/	AE (two DenseNet encoders), LF: MSE loss and cross-entropy, train 60% val 20% test 20%	**Accuracy:** 98.39 **Precision:** 77.73 **Recall:** 77.24 **Specificity:** 99.16 **F1 score:** 76.96
Ding et al. [55]	MfTransNet	Classification of cardiac DE-MRI	EMIDEC Challenge dataset [56]	MRI	Spatial augmentation	Transformer, LF: CIoU loss, train 70% test 30%	**Accuracy:** 84.97%
Guo et al. [57]	+DLKC/nDLKC	LV, RV and Myo segmentation	ACDC [58], UK Biobank	MRI	/	CNN: U-net and Isensee2017, LF: Dice loss and cross-entropy	**ACDC (DSC)** LV: 1.7, Myo: 1.3, RV: 9.7 **UKBB (DSC)** LV: 2.7, Myo: 2.8, RV: 4.6

Continued on next page

Reference	Approach Name	Application	Dataset	Imaging Modality	Data Preprocessing	DL Technique and Training Strategy	Test Performance
Yu et al. [59]	PABVS	LV and RV segmentation	LV Full Quantification Challenge of MICCAI 2018 [60]	MRI	landmark labelling, rotation, ROI cropping, resizing 80×80	DenseNet and CNN, train 69% test 31%	**DSC** EpiLV: 0.915 ± 0.118 EndoLV: 0.871 ± 0.110 RV: 0.843 ± 0.080
Xian et al. [61]	DAG-Net	Segmentation of AA, LAC, LVC, RVC, Myo	MM-WHS Challenge dataset [62], MS-CMRSeg dataset [63]	MRI, CT	Data augmentation (random rotation, elastic deformation)	CNN and FCSA, RSA, train 80% test 20%	**MM-WHS (DSC)** CT→MRI: 77.39 ± 7.87 MRI→CT: 84.60 ± 4.31
Chakravarty and Sivaswamy [64]	RACE-net	Segmentation of LA	LASC [65]	MRI	Crop ROI, data augmentation (random translations)	RNN, LF: weighted sum of intermediate dice coefficients, train 33.33% test 66.66%	**DSC:** 0.91 / 0.04
Pace et al. [66]	/	Whole heart segmentation	HVSMR challenge dataset [67]	MRI	Data augmentation (random affine and nonlinear transformations, flips, intensity shifts, Gaussian noise)	RNN, LF: proposed loss function, 4-fold cross validation	**DSC:** 92.6 ± 3.0
Khened et al. [68]	DFCN-C	LV, RV, Myo segmentation	ACDC, LVSC	MRI	Data augmentation, Capture ROI, Resizing 256×256	DenseNet and FCN, LF: Dice loss and weighted cross-entropy, train 66.66% test 33.33%	**DSC:** 0.91 (0.04) **HD** 5.43 (4.40)
Yuwen et al. [69]	GANSA	Whole heart segmentation	MM-WHS	CT, MRI	/	GAN with self-attention mechanism, LF: cyclic consistency loss	/
Chen et al. [70]	TransUNet	LV, RV, Myo segmentation	ACDC	MRI	Resize to 224×224	Transformer, FCN (U-Net), train 70% test 20% val 10%	**DSC:** 77.48 **HD:** 31.69

Continued on next page

Reference	Approach Name	Application	Dataset	Imaging Modality	Data Preprocessing	DL Technique and Training Strategy	Test Performance
Kou et al. [71]	DenseBiasNet	Segmentation of myocardium and blood pool	CMRxMotion-2022 dataset [72]	MRI	/	FCN with dense bias connections and VAE, LF: combined cross-entropy, Dice loss and reconstruction loss, train 80% val 20%	**DSC:** LV: 0.82, RV: 0.68, Myo: 0.64 **HD:** LV: 15.21, RV: 21.16, Myo: 17.532
Cui et al. [73]	GBCUDA	Whole heart segmentation	MM-WHS	CT, MRI	/	GAN with self-attention mechanism, spectrum normalization, LF: knowledge distillation loss, train 33.33% test 66.66%	**DSC:** MRI→CT: 81.5 CT→MRI: 59.2 **ASD:** MRI→CT: 5.8 CT→MRI: 4.9
Decourt and Duong [74]	DT-GAN	Pericardial membranes segmentation	Private	MRI	Resize to 256×256, Apply CLAHE equalizer	GAN (segmentation + discriminator network), LF: adversarial and distance map loss, Cross validation	**DSC:** Endo: 0.87 ± 0.11 Epi: 0.94 ± 0.08 Myo: 0.81 ± 0.08
Xu et al. [75]	LeViT-UNet	LV, RV, Myo segmentation	ACDC	MRI	Data augmentation (random flipping, rotations), resize to 224×224	Transformer, FCN (U-Net), transfer learning, LF: cross-entropy and Dice loss, train 80% test 20%	**DSC:** 90.32 **HD:** 16.84
Fu et al. [76]	TF-UNet	LV, RV, Myo segmentation	ACDC	MRI	Data augmentation (rotation, scaling, Gaussian blur, noise, brightness), crop	Transformer, FCN (U-Net), transfer learning, LF: cross-entropy and Dice loss, train 70% test 20% val 10%	**DSC:** 91.72
Dezaki et al. [77]	/	Detection of cardiac phases	Private	ECHO	Crop ROI, resize 120×120	DenseNet and GRU-RNN, LF: global extrema structure loss, train 60% val 20% test 20%	**Average Error Prediction** ED: 0.20 ± 0.76 ES: 1.43 ± 1.3

Continued on next page

Reference	Approach Name	Application	Dataset	Imaging Modality	Data Preprocessing	DL Technique and Training Strategy	Test Performance
Reynaud et al. [78]	/	Detection of cardiac phases	Echonet-Dynamic dataset [79]	ECHO	Pad all frames with zeros from 112×112 to 128×128	Transformer (Residual AE network + BERT model), LF: MSE loss and weighted cross-entropy, train 75% val 12.5% test 12.5%	**Average Frame Difference** ED: 2.86 (6.43) ES: 7.88 (11.03)
Chen et al. [80]	Res-CRNN	Real-time cardiac cine reconstruction	Private	MRI	Undersampling	Residual convolutional RNN, LF: MSE and SSIM	**CSM:** 3.84 (0.33) **RT:** 25.70 (2.68)
Sarvari and Sridevi [81]	EBRSA-bi LSTM	Reconstruction of CT images	COVID-CT dataset [82], SARS CoV-2 CT-scan	CT	Transform to grayscale, remove noise/distortions, ECF preprocessing	RNN, self-attention, LF: proposed loss function, train 80% val 10% test 10%	**RT:** 0.085 **PSNR** Dataset 1: 45.152 Dataset 2: 46.123 **RMSE** Dataset 1: 0.0026 Dataset 2: 0.0022
Zhao et al. [83]	LSRGAN	Ultrasound image generation	Private	MRI	Data augmentation (horizontal/vertical flips, random cropping)	GAN, LF: least squares	/
Skandarani et al. [84]	VAE-GAN	Realistic CMR images generation	ACDC, Sunnybrook [85]	MRI	Data augmentation (rotations, flips, shifts)	VAE, GAN and SPADE, ACDC: train 64% test 36%, Sunnybrook: train 67% test 33%	/
Tiago et al. [86]	/	Synthetic 3D cardiovascular ultrasound image generation	Private	ECHO	Data augmentation (blurring, rotation)	GAN (Pix2pix model), LF: L1 loss and conditional GAN loss, 5-fold cross validation	**DSC:** 0.844 ± 0.047
Lyu et al. [87]	/	Cardiac cine MRI reconstruction	Private	MRI	/	Transformer-based GAN, LF: L1-pixel loss, VGG-16 perceptual loss and adversarial loss, train 80% test 20%	**PSNR:** 34.58 (0.57) **SSIM:** 0.8745 (0.0075) **RMSE:** 1.89 (0.10)

Continued on next page

Reference	Approach Name	Application	Dataset	Imaging Modality	Data Preprocessing	DL Technique and Training Strategy	Test Performance
Yoon et al. [88]	SADM	Generation of longitudinal medical images	ACDC	MRI	Resize to 128×128×32	Transformer, train 66.66% test 33.33%	**SSIM:** 0.851 **PNSR:** 28.992 **NMRSE:** 0.153
Zeleznik et al. [89]	/	Quantification of coronary calcium	FHS [90], NLST [91], ROMICAT–II [92]	CT	Data augmentation (rotation, translation)	CNN, FHS: train 71% test 29%, NLST and ROMICAT-II used for testing	/
Liu et al. [93]	/	X-ray radiology report generation	Open-I [94], MIMIC-CXR [95]	X-ray	/	RNN, reinforcement learning, LF: cross-entropy loss	**Accuracy** Open-I: 0.867 MIMIC-CXR: 0.918
Chang et al. [96]	MC-Net	Assessment and correction of CMR images	MM-WHS, LVSC	CT, MRI	Resize to 512×512	RNN, LF: intersection distance loss, train 80% test 20%	**Mean Absolute Distance:** 0.499 **DSC:** 0.994
Bello et al. [97]	/	Patients' survival prediction with pulmonary hypertension	Private	MRI	Automatic segmentation of ventricles, image registration	FCN and DAE, LF: Cox partial likelihood loss, train 70% test 30%, Bootstrap validation, 6-fold cross validation	/
Beetz et al. [98]	/	Prediction of ED outputs from ES inputs and ES outputs from ED inputs	UK Biobank	MRI	/	Extended version of PointNet++, Train 80% test 20%	**Chamfer distance:** 1.66 ± 0.62

MCD: Monte Carlo Dropout, **LF**: Loss Function, **IVUS**: Intravascular Ultrasound, **DSC**: Dice Score, **HD**: Hausdorff Distance, **VAE**: Variational AE, **ASD**: Average Surface Distance, **ECHO**: Echocardiogram, **CSM**: Coil Sensitivity Maps, **RT**: Reconstruction Time, **MSE**: Mean Square Error, **SSIM**: Structural Similarity Metric, **RMSE**: Root Mean Square Error, **PSNR**: Peak Signal-to-Noise Ratio, **NMRSE**: Normalized Root Mean Square Deviation.

The reviewed studies indicate that DL techniques are widely used across various cardiac image processing and analysis tasks, enhancing both accuracy and automation. CNNs are among the most commonly applied DL architectures in this field primarily due to their ability to autonomously extract distinctive features from input images [57], [59], [61], [49], [51], [89], [53]. Specifically, CNNs can accurately outline the boundaries and contours of various cardiac structures. They have been utilized in a range of cardiac image analysis tasks, including the segmentation of key structures, such as the LV, RV, myocardium, and whole heart; vessel categorization; ejection fraction assessment; and diagnosing various heart conditions. CNNs have proven effective in cardiacstructures segmentation [61], reconstructing cardiac MRI images [80], and quantifying coronary calcium [89], among other applications. The Fully Convolutional Network (FCN), a specific variant of the CNN architecture, is primarily used for image segmentation due to its unique structure, which consists exclusively of convolutional layers. FCNs are used either on their own [97] or in combination with other DL architectures [68], [71].

RNNs have also been applied in cardiac image processing, particularly for handling sequential data, such as ECGs and time-series cardiac imaging data. RNNs are well-suited for capturing temporal relationships and patterns in data, making them highly effective for analyzing ECG signals in arrhythmia detection and classification. Additionally, RNNs have been used in the segmentation of cardiac images, including left atrium segmentation and MI detection [64], [66], as well as cardiac phase identification [77], and reconstruction tasks for cardiac imaging such as MRI [80] and CT scans [81]. They have also been utilized in automated report generation for X-ray radiography [93]. Moreover, RNNs have demonstrated effectiveness in predicting clinical outcomes, such as left ventricular function and quantification and analyzing and correcting CMR images [96] using longitudinal cardiac imaging data. AEs have also been employed in cardiac image analysis for various purposes. Autoencoders are neural networks designed to learn data compression and subsequent reconstruction. They have been used for tasks such as myocardium and blood pool segmentation [71], classification of pediatric cardiac ultrasound images [54], cardiac phase identification [78], 3D cine MRI reconstruction [84], and survival prediction in patients with pulmonary hypertension [97].

Recently, significant progress in DL methods has been achieved within cardiac imaging, demonstrating strong potential to enhance the identification and treatment of CVDs. These advancements are largely due to the rise of advanced DL frameworks, including GANs and Transformers, which have exhibited noteworthy performance across diverse cardiac imaging applications. For instance, GANs have been applied to several tasks like image creation [83], [84], [86], and [87], as well as segmentation and denoising, with an emphasis on image dataset preprocessing [73], [74], [69]. In tasks involving image synthesis and production, GANs can generate realistic cardiac images, which supplement the often-limited existing data. Furthermore, GANs have yielded impressive results in

segmenting distinct heart structures. GANs are increasingly favored in cardiac image analysis because they can learn complex and nonlinear mappings between input and output, which enhances the accuracy of heart structure segmentation. The availability and popularity of image-to-image translation models have further supported the application of GANs for cardiac image segmentation and reconstruction.

Unlike GANs, which mainly focus on segmentation and reconstruction, Transformers are utilized across all of the cardiac imaging tasks, such as segmentation, classification, detection, reconstruction, and prediction. Transformers have grown in popularity in cardiac imaging due to their ability to identify long-range dependencies and represent key features effectively. These models analyze a sequence of image patches, enabling them to replace traditional convolutions in deep neural networks. Consequently, Transformers have been employed in a range of tasks, including segmentation of the LV, RV, and myocardium [70], [75], [76], classification of cardiac MRI images [55], detection of cardiac phases [78], reconstruction of cardiac cine MRI images [87], [88], and predicting left ventricular ejection fraction [78]. The attention mechanisms in Transformers allow for a deep understanding of complex spatial relationships within images, making them ideal for tasks that require contextual spatial comprehension and recognition of long-term dependencies.

The reviewed studies indicate that MRI is the most commonly used imaging modality in DL-based approaches for CAD in CVD diagnosis. This focus on MRI can be largely attributed to the availability of several high-quality public datasets, such as ACDC and the UK Biobank, which have been highly valuable in advancing MRI research for cardiac imaging. These datasets provide comprehensive, standardized data, making MRI a preferred choice for researchers. In contrast, other modalities, such as nuclear cardiology and echocardiography, are less frequently used, primarily due to the limited access to similarly extensive and publicly available datasets. MRI's widespread use in these studies stems not only from data availability but also from its ability to capture high-resolution, detailed images of cardiac structures, which is crucial for tasks like segmentation, classification, and disease detection. This precision in imaging enables the development of accurate DL models designed for various cardiac analysis tasks, further emphasizing MRI's role as the imaging modality of choice in this domain.

To tackle data-related challenges, researchers have employed a variety of preprocessing techniques to prepare data before feeding it into DL models. These preprocessing steps, including data normalization, scaling, filtering, feature extraction, and data augmentation, are designed to reduce image noise, improve contrast, and standardize intensity levels across images [59], [61]. Among these, Region of Interest (ROI) selection is also widely used [64], enabling researchers to focus on specific areas of interest, such as particular heart structures or vessels, thereby excluding extraneous details that may impede the model's accuracy and efficiency. The majority of the studies reviewed employed

at least one preprocessing step, with some specifically implementing ROI extraction to enhance focus on relevant cardiac structures. Methods for extracting ROI vary, with certain studies, such as [59] and [64], opting to crop images to predefined dimensions for targeted analysis. By incorporating these preprocessing techniques, researchers have developed more robust and effective DL models for tasks such as segmentation, classification, and disease detection in cardiac imaging. Moreover, the models in studies summarized in Table 2.1 consistently demonstrate improved evaluation metrics over those lacking a preprocessing phase, underscoring the positive impact of these techniques on model performance.

2.4 Conclusion

This chapter has provided a comprehensive examination of the cardiovascular system, CVDs, and the diverse diagnostic techniques employed in cardiology. It began with an in-depth analysis of the heart's anatomy and physiology, followed by a detailed investigation of the cardiac cycle and its phases. Key CVDs were then classified and defined, accompanied by an epidemiological overview that underscored the prevalence, incidence, mortality, and morbidity rates associated with these conditions, thus highlighting their profound global health impact.

The chapter also explored various diagnostic approaches, from clinical evaluations to advanced imaging methods, with particular emphasis on CMR due to its non-invasive nature and ability to provide detailed insights into cardiac anatomy and function, underscoring its value in modern diagnostics. In addition, the chapter discussed the transformative role of CAD systems and AI in cardiology. Advanced DL techniques, such as CNNs and ViTs, have shown promising applications in this field. The subsequent chapter will provide a more detailed examination of these techniques, exploring their applications in image segmentation.

Chapter 3

Deep Learning Techniques for Medical Image Analysis

3.1 Introduction

In recent years, there have been significant advances in medical imaging, particularly in cardiac image segmentation, due to the advent of DL techniques. This chapter examines the general process of image-based CAD systems and explores how DL is transforming medical image analysis. By addressing fundamental concepts, architectures, and optimization strategies, this chapter provides a comprehensive overview of the role of DL in improving the accuracy and efficiency of image-based CAD systems.

We begin by presenting the general process of an image-based CAD system, which includes stages such as preprocessing, assisted diagnosis (including classification and segmentation), and decision-making. We then provide an in-depth examination of key DL techniques for medical image analysis, including CNNs and ViTs. We analyze their basic components, architectures, and applications in medical image segmentation. Finally, we discuss optimization strategies for DL models, focusing on TL, which improves the performance and applicability of DL techniques in medical imaging.

This chapter is organized as follows: Section 3.2 discusses the general process of an image-based CAD system, covering its main stages. Section 3.3 provides an overview of DL techniques, focusing on the basic concepts of CNNs and ViTs and the optimization strategies for DL models.

3.2 General Process of an Image-Based CAD System

The general process of an imaging-based CAD system consists of several interconnected phases designed to analyze medical images and provide diagnostic support. The first and most important step is the collection and creation of the dataset, which serves as the foundation for the entire system. This dataset is processed and refined in the subsequent *Preprocessing* phase to prepare it for further analysis.

Once the data is preprocessed, the system enters the *Aided Diagnosis* phase, where the images are analyzed to extract meaningful patterns and features. This phase typically involves tasks such as *Classification* and *Segmentation*, which contribute to understanding the data and identifying relevant diagnostic insights. The outcomes of these tasks are then utilized in the final step, where the system produces a *Decision* to assist clinicians in their diagnostic process. This structured process enables CAD systems to integrate advanced computational techniques into medical imaging workflows, enhancing the accuracy, efficiency, and reliability of diagnostic decision-making. Figure 3.1 illustrates the general process of an image-based CAD system, highlighting its key phases and their interconnections.

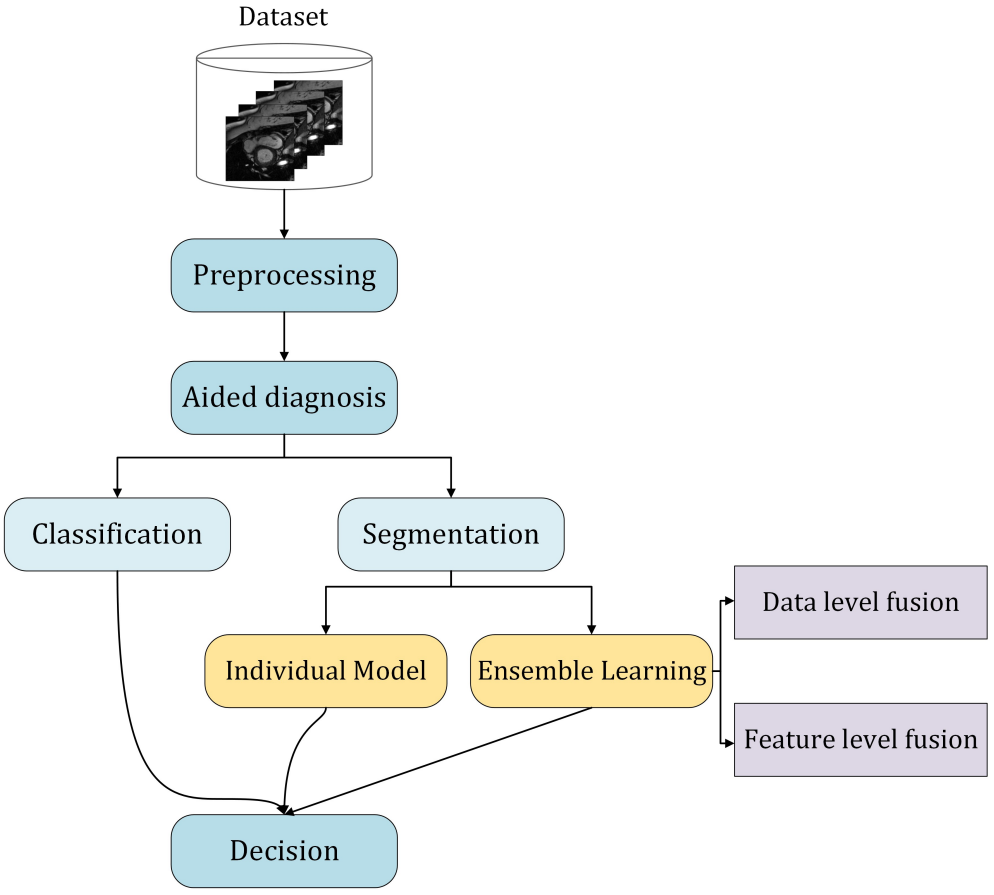

FIGURE 3.1: General process of an imaging-based CAD system.

3.2.1 Preprocessing

The preprocessing phase in an image-based CAD system is a crucial step in preparing raw medical imaging data for accurate and reliable analysis. Medical images often contain artifacts, noise, or inconsistencies resulting from the acquisition process, which can hinder the performance of subsequent diagnostic tasks [99]. Preprocessing addresses these issues through various techniques aimed at enhancing the quality and standardizing the data.

The key objectives of the preprocessing phase include noise reduction, image enhancement, normalization, and artifact removal [100]. Noise reduction methods, such as Gaussian filtering or median filtering, are used to suppress random variations in image intensity, ensuring a clearer representation of anatomical structures [101]. Image enhancement techniques, such as contrast stretching or histogram equalization, improve the visibility of critical features in the image [102]. Normalization is applied to standardize the intensity values across the dataset, ensuring consistency, especially when images are acquired from different modalities or devices [103]. Additionally, artifact removal eliminates unwanted distortions, such as scanner-induced artifacts, that may obscure relevant information.

In some cases, preprocessing may also involve resizing or cropping the images to focus on regions of interest (ROI) and reduce computational complexity. Furthermore, for datasets involving multimodal imaging, image registration techniques are used to align images from different sources, ensuring that corresponding anatomical structures are accurately matched [104].

3.2.2 Assisted Diagnosis

The assisted diagnosis phase is the central analytical step in an image-based CAD system, where computational techniques are used to process medical images and derive meaningful diagnostic insights. This phase transforms preprocessed data into practical information by identifying patterns, abnormalities, or structures that are relevant to a clinical diagnosis.

The principal objective of the assisted diagnosis phase is to automate or augment the identification of diseases or pathological features, making use of advanced algorithms to enhance accuracy and reduce the subjectivity inherent in manual diagnosis. The main tasks in this phase include the classification of images into diagnostic categories and the segmentation of ROIs such as lesions, tumors, or anatomical structures. These tasks often rely on ML and DL models, which are capable of handling the complex and high-dimensional data associated with medical imaging.

3.2.2.1 Classification

The classification task involves assigning an input image x to one of several predefined categories or classes $\{C_1, C_2, \ldots, C_n\}$. The objective is to learn a mapping function $f : X \to C$, where X is the input space (e.g., pixel intensities or features extracted from the image), and C is the set of output classes. This mapping is achieved by modeling the conditional probability distribution $P(C_i \mid x)$ for each class C_i, such that:

$$\hat{y} = \arg\max_{C_i \in C} P(C_i \mid x), \tag{3.1}$$

where $\hat{y}$ is the predicted class. The function $P(C_i \mid x)$ is often approximated using ML models.

Classification tasks have been extensively applied in cardiac imaging to identify and diagnose cardiovascular conditions using modalities such as echocardiography, cardiac MRI, and CT scans. These systems focus on detecting arrhythmias, MI, structural heart diseases, and other abnormalities by leveraging ML and DL techniques [49, 55]. In recent advancements, DL techniques have been applied to cardiac imaging for the detection and classification of arrhythmias. Lu et al. [105] introduced a DL-based reconstruction method for cardiac MRI, enhancing image quality and facilitating more accurate speckle-tracking echocardiography. Classification algorithms have also been employed to detect infarcted myocardial tissue. For instance, Chen et al. [106] developed a DL model that automatically identifies MI by analyzing late gadolinium enhancement images, achieving high accuracy in distinguishing infarcted from non-infarcted myocardium. ECGs have also been classified using DL methods for the detection of cardiac abnormalities and COVID-19-related impacts. Islam et al. [49] developed an improved CNN to classify ECG-derived images for the detection of covid-19 related impacts. Moreover, coronary CT angiography has been used in CAD systems to classify the presence and severity of coronary artery disease. Advanced models, such as 3D CNNs, have been applied to volumetric coronary CT angiography data for the automatic detection and grading of stenosis [53]. These studies collectively highlight the effectiveness of classification tasks in cardiac imaging, enabling automated, accurate, and reproducible diagnoses across a wide range of cardiovascular conditions.

3.2.2.2 Segmentation

Segmentation in CAD systems is the process of partitioning an image into distinct, meaningful regions to facilitate analysis and interpretation. This process is critical in medical

imaging, where identifying specific anatomical structures or pathological regions is essential for diagnosis, treatment planning, and disease monitoring. Two principal types of segmentation are widely employed: *semantic segmentation* and *instance segmentation*.

Semantic Segmentation

Semantic segmentation assigns a class label to each pixel in the image, categorizing every region into predefined classes. The objective is to create a pixel-wise classification where each pixel belongs to a specific category, such as "tumor," "healthy tissue," or "background." Formally, for an image domain I and a set of classes L, semantic segmentation aims to map each pixel $i \in I$ to a class label $l \in L$:

$$S(i) = l, \quad \forall i \in I. \tag{3.2}$$

This approach does not differentiate between individual instances of the same class but focuses on the overall class distribution within the image. Semantic segmentation is frequently used in medical imaging for tasks like delineating organ boundaries or identifying regions affected by disease. DL architectures such as FCNs are commonly applied to perform semantic segmentation due to their ability to learn spatial hierarchies and produce detailed pixel-wise predictions.

Instance Segmentation

Instance segmentation extends the concept of semantic segmentation by not only labeling each pixel according to its class but also distinguishing between individual objects or regions within the same class. The goal is to provide a unique identifier for each instance of a class in addition to its label. For an image domain I, a set of classes L, and instances j, instance segmentation can be expressed as:

$$S(i) = \{(l, \text{instance}_j)\}, \quad \forall i \in I. \tag{3.3}$$

This capability is particularly important in scenarios where multiple objects of the same type need to be individually analyzed, such as counting or measuring multiple lesions, tumors, or anatomical structures in a medical image. Instance segmentation combines object detection and pixel-level segmentation, often using frameworks such as Mask R-CNN or hybrid transformer-based models to achieve accurate results.

In cardiac image segmentation, individual models and ensemble learning are two prominent approaches. Individual models, such as U-Net [107], have been widely used due to

their ability to capture intricate anatomical details in cardiac MRI images [108]. However, depending only on a single model may not fully capture the variability present in cardiac images. To address this, ensemble learning combines multiple models to improve segmentation accuracy and robustness. For instance, Dang et al. [109] proposed a two-layer ensemble of DL models for medical image segmentation, demonstrating enhanced performance compared to individual models [110]. By integrating diverse models, ensemble learning leverages their complementary strengths, leading to more reliable cardiac image segmentation.

Ensemble approaches in cardiac image segmentation often employ data-level fusion and feature-level fusion techniques. Data-level fusion involves combining inputs from multiple imaging modalities or different views before processing. For instance, Wu et al. [111] introduced a fusion-attention mechanism for cardiac MRI image segmentation, effectively integrating information from various sources to improve segmentation outcomes. On the other hand, feature-level fusion combines features extracted from different models or modalities during the processing stage. An example is the use of a dilated convolution network with an edge fusion block and directional feature maps for cardiac MRI segmentation, which integrates multi-scale features to enhance boundary delineation [112]. Both fusion strategies aim to exploit complementary information, thereby enhancing the accuracy and reliability of cardiac image segmentation.

Evaluation Metrics for Image Segmentation

Common metrics for assessing the performance of a segmentation model include the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), and Intersection over Union (IoU) [113]. Suppose that G and P are used to represent the ground truth and prediction results of image segmentation, respectively.

The Dice Similarity Coefficient (DSC). Dsc is a widely used metric in medical image segmentation to evaluate the overlap between two regions of interest: the segmented region P and the ground truth G. It is specifically designed to measure the spatial agreement between P and G, offering a quantitative evaluation of their intersection relative to their combined volume or surface. The DSC is defined as:

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

Here, $|G \cap P|$ denotes the number of voxels (or pixels) common to G and P, while $|G|$ and $|P|$ represent the total number of voxels/pixels in G and P, respectively. The DSC

ranges from 0 to 1, where a value of 1 indicates perfect overlap, and 0 indicates no overlap.

The Jaccard Distance. Often referred to in the context of image segmentation as Intersection over Union (IoU), is a metric used to quantify the similarity between two regions, such as a predicted segmentation P and the ground truth G. It evaluates the proportion of overlap between the predicted and actual regions relative to their combined area. The IoU is defined as:

$$\text{IoU}(G, P) = \frac{|G \cap P|}{|G \cup P|} \tag{3.5}$$

Here, $|G \cup P|$ denotes the total number of voxels/pixels in either G or P, encompassing both their union and any non-overlapping areas. The IoU ranges from 0 to 1, where a value of 1 indicates perfect overlap, and 0 indicates no intersection between the regions.

The Hausdorff Distance (HD). HD is a metric used in image segmentation to measure the extent of mismatch between the boundaries of a predicted segmentation P and the ground truth G. It quantifies the maximum distance between the closest points of these two sets, essentially capturing the largest deviation between corresponding boundary points. The HD is defined as:

$$\text{HD}(G, P) = \max\left\{ \max_{g \in G} \min_{p \in P} \|g - p\|_2, \max_{p \in P} \min_{g \in G} \|g - p\|_2 \right\} \tag{3.6}$$

In this formula, $\|g - p\|_2$ represents the Euclidean distance between a point g in the ground truth G and a point p segmentation segmentation P. The expression $\min_{p \in P} \|g - p\|_2$ finds the closest point in P to a given point in G, while $\max_{g \in G}$ identifies the point in G with the maximum of these minimum distances. The formula is symmetric, also considering the maximum of the minimum distances from P to G. This distance effectively measures the worst-case scenario of boundary alignment, with a lower HD indicating a closer match between P and G boundaries.

The Average Symmetric Surface Distance (ASSD). ASSD is a metric used in image segmentation to measure the average distance between the surfaces of the predicted segmentation P and the ground truth G. Unlike the HD, which captures the maximum boundary discrepancy, the ASSD provides a more balanced assessment by averaging the surface distances in both directions. The ASSD is mathematically defined as:

$$\text{ASSD} = \frac{\sum_{g \in G} d(g, P) + \sum_{p \in P} d(p, G)}{|G| + |P|} \tag{3.7}$$

In this formula, $d(g, P)$ represents the shortest Euclidean distance from a point g on the ground truth surface G to the closest point on the predicted surface P, and $d(p, G)$ is the shortest distance from a point p on the predicted surface P to the ground truth

G. The sums $\sum_{g \in G} d(g, P)$ and $\sum_{p \in P} d(p, G)$ calculate the total of these distances for all points on each surface. Finally, $|G|$ and $|P|$ are the number of points on G and P, respectively.

Loss Functions for Image Segmentation

In image analysis tasks, improving model accuracy relies not only on the architecture but also significantly on the choice of loss functions [114]. The loss function plays a crucial role by computing the total error during the training process and adjusting the model's weights through backpropagation. Various loss functions have been developed to address different challenges across domains, with some adaptations of existing ones. The most commonly used loss functions in segmentation tasks are Cross-Entropy loss, Dice loss, or a combination of these two functions, offering a robust approach to optimizing model performance [115, 116].

Cross-Entropy loss. The Cross-Entropy Loss function is a widely used metric in classification and segmentation tasks, particularly when dealing with multi-class problems. It measures the difference between the true labels and the predicted probabilities, aiming to minimize the divergence. The Cross-Entropy Loss is defined by:

$$L_{CE}(G, P) = -\sum_{i=1}^{N} G_i \log(P_i) \tag{3.8}$$

This equation calculates the average loss for each pixel in an image, where G_i represents the true probability for the i^{th} class, and P_i denotes the predicted probability for the same class. This approach helps the model generate probability maps that closely match the actual segmentation masks while penalizing inaccurate predictions more heavily. By minimizing the cross-entropy loss during training, the model improves its precision in image segmentation.

Despite its widespread use, this method can be biased by dataset imbalances, particularly in medical image segmentation, where the majority class often dominates. Specifically, the large proportion of a medical image may consist of the default background label, which can lead to issues in accurately segmenting the specific organ of interest. To address this, a weighted cross-entropy loss is introduced, as described by Ronneberger et al. [107], given by:

$$L_{WCE}(G, P) = -\sum_{i=1}^{N} G_i w_i \log(P_i) \tag{3.9}$$

In this formula, w_i is a weight factor for the i^{th} class, added to adjust the standard cross-entropy loss equation. However, this adjustment does not fully resolve the issue,

as cross-entropy loss calculates the average per-pixel loss without accounting for the relationships between adjacent pixels, which can include boundaries.

Dice loss. Dice loss is a widely used metric in computer vision for assessing the similarity between two images. It builds on the Dsc, which was later adapted as the Dice loss function for segmentation tasks. This approach was first applied in the field of medical image segmentation by Milletari et al. [117]. The Dice loss is defined as:

$$L_{\text{Dice}}(G, P) = 1 - \frac{2\sum_{i=1}^{N} G_i P_i}{\sum_{i=1}^{N} G_i + \sum_{i=1}^{N} P_i + \epsilon} \tag{3.10}$$

In this formula, the summation is performed over N pixels, with ϵ being a small constant added to prevent division by zero. The Dice coefficient evaluates the overlap between G and P, yielding a score between 0 and 1, where 1 indicates perfect overlap. By considering pixel-level information in both global and local contexts, Dice loss often provides more accurate results compared to cross-entropy loss. However, this loss could be unstable [118] and is often mixed with the cross-entropy loss [119].

Combined Cross-Entropy Dice loss. The combined loss function leverages the strengths of both cross-entropy loss and Dice loss to address challenges in imbalanced datasets. The cross-entropy loss focuses on improving pixel-wise classification accuracy, effectively guiding the model to distinguish between different classes. Meanwhile, the Dice loss is designed to enhance segmentation performance by ensuring that the model accurately captures and respects the structure of the target regions, even when these regions are small or underrepresented. The combined loss function is defined by:

$$L_{\text{comb}} = L_{CE} + L_{\text{Dice}} \tag{3.11}$$

3.2.3 Decision

The decision phase is the final step in a CAD system, where outputs from classification and segmentation tasks are combined into clinically actionable conclusions. This phase directly interfaces with healthcare professionals, translating computational results into diagnostic insights [120]. Classification outputs, such as disease labels or probability scores, help determine the presence, type, or stage of a condition, while segmentation outputs provide spatial and volumetric data crucial for precise assessment and treatment planning.

In clinical practice, these outputs support decision-making by offering detailed and reproducible information. Classification results can highlight high-risk cases for prioritization, and segmentation maps precisely identify affected regions, aiding interventions.

To ensure clinical usability, outputs must be interpretable, with tools like overlaid segmentation maps or highlighted regions to facilitate clinician verification. Robustness to variations in imaging conditions and patient demographics is also essential for consistent performance across diverse environments.

3.3 Deep Learning Techniques for Medical Image Analysis

3.3.1 Deep Learning Concepts

DL, a subset of ML, offers significant advantages over traditional ML techniques, particularly in its ability to automatically learn features from data [47]. Traditional ML often requires manual feature extraction, necessitating the involvement of domain experts in determining which features are pertinent to a given task. This process can be challenging and may not always yield the best results.
In contrast, DL employs representation learning, where algorithms learn both the representation and the mapping from representation to output. This approach leads to superior performance as it eliminates the need for hand-crafted features [121]. DL models utilize artificial neural networks (ANNs) with multiple layers, allowing them to learn complex patterns and representations from the data. These ANNs, inspired by the structure of the human brain, consist of interconnected neurons that process and transmit information. The deep layers within these ANNs serve as feature extractors, automatically generating representations that are not specific to a particular task but are generally applicable across various applications [47]. A basic example of an ANN is a neuron or perceptron, which can be described as:

$$f_n(x) = g(W \cdot x) + B \tag{3.12}$$

Here, the input is represented by x, the non-linear activation function by g, the weight and bias by W and B, respectively, and the dot product operation by $\cdot$. In order to approximate the relationship between the input x and the intended output, the neuron tries to choose the best weight and bias combination. However, a neuron's linear separability places limitations on its solution space. The multi-layer perceptron (MLP), which links several perceptrons to map input values to output values, was created to overcome this constraint. An MLP consists of many neurons organized into one or more non-linear layers, known as hidden layers. A simple example of an MLP with a single hidden layer can be described as:

$$f(x) = g_2\big(W_2 \cdot g_1(W_1 \cdot x + B_1) + B_2\big) \tag{3.13}$$

Where the activation function, weight, and bias of the i^{th} layer are represented, respectively, by g_i, W_i, and B_i.

The predictive ability of ANNs is derived from the multi-layered architecture of MLPs. Figure 3.2 shows an MLP with a single hidden layer consisting of an input layer with three neurons, a hidden layer with two neurons, and an output layer with one neuron. Each neuron in the hidden layer processes the values from the input layer using a weighted combination, followed by a non-linear activation function. Every neuron in this MLP is fully connected to all neurons in the subsequent layer. MLPs can only be used with one-dimensional training sets, though. Neural network performance can be improved by applying convolution operations to better capture higher-dimensional patterns, such as edges and contours.

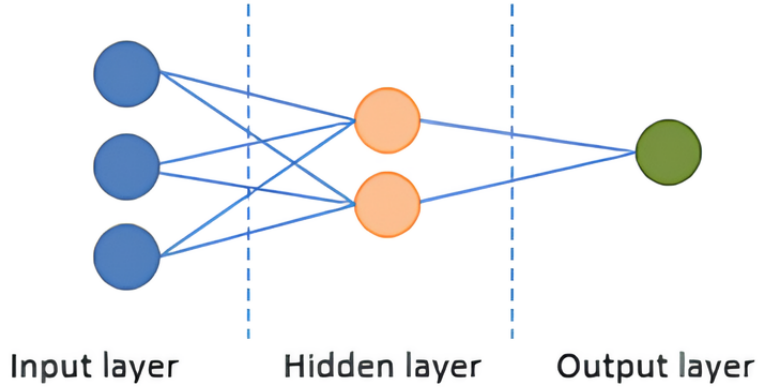

FIGURE 3.2: Basic MLP with a single hidden layer.

3.3.2 Convolutional Neural Networks

CNNs, proposed by LeCun et al. [47], are deep neural networks that treat input images by dragging filters across the data. The CNN structure is inspired by the visual cortex of the animal brain [122]. In contrast to traditional neural networks, which only contain a classification part (Figure 3.2), the architecture of the CNN has an upstream convolutional part and therefore has two very distinct parts: (a) A convolutive part that aims to extract characteristics specific to each image by compressing them to reduce their initial size. In short, the input image passes through a series of filters, creating new images called feature maps. Finally, the resulting feature maps are concatenated into a characteristic vector called the CNN code. (b) A classification part where the CNN code obtained at the output of the convolutive part is provided as input in a second part, consisting of fully connected layers (MLP). The role of this part is to combine the characteristics of the CNN code to classify the image.

3.3.2.1 Fundamental Components of CNN

CNN model consists of three key layers for classifying images: convolution, pooling, and fully connected layers. The convolution and pooling layers are in charge of extracting features (convolutive part), while the fully connected layer aims to classify the input data (classification part). As the name of the model implies, the convolution layer is the most important; It reduces the size of the input image by extracting features such as edges, colors, texture, and gradient orientation. To do that, the convolution layer uses learnable filters to perform convolution operations on the input image. As shown in Figure 3.3, the result of these operations is a tensor called an output feature map. The nonlinear activation functions act as triggers and receive the output data from the convolution operations. Activation functions come in different forms, including sigmoid, SoftMax, tangent (tanh), and the rectified linear unit (ReLU).

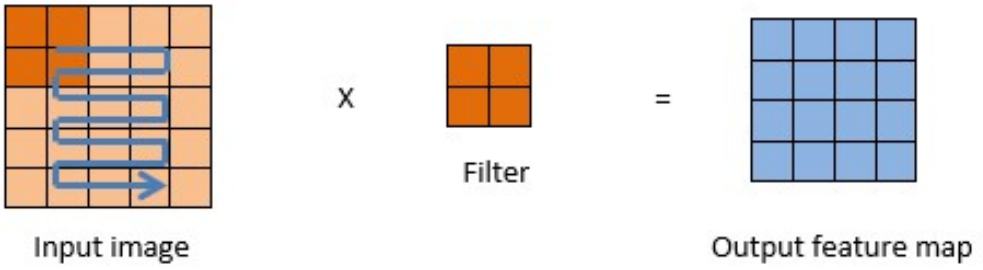

FIGURE 3.3: Convolutional operation.

The pooling layer aims to decrease the computational power required to process the data by reducing dimensionality. It performs a non-linear down-sampling of the convolved feature map (Figure 3.4). There are multiple forms of pooling layers, such as max pooling and average pooling layers. The input is divided into several rectangular patches resulting from the pooling operation. Depending on the pooling type selected, every patch is replaced by a single value (i.e., the maximum or average value of the rectangular patch).

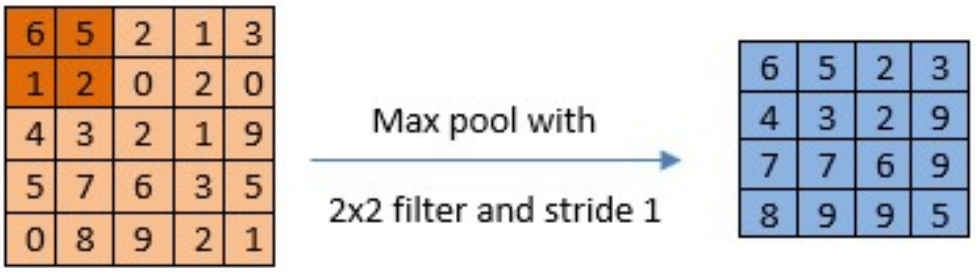

FIGURE 3.4: Max-pooling operation.

The last layers of CNNs are fully connected layers. These layers are similar to traditional neural networks. The feature maps are flattened before being passed to the fully connected layer to prevent a dimension mismatch between the outputs of the convolution/pooling layers and the fully connected layer. The feature maps' dimensions are reshaped into a vector shape during this flattening operation (one-dimensional). The last layer of the fully connected layers is in charge of classification.

3.3.2.2 Overview of CNN Architectures

The first proposed CNN architecture by LeCun et al. [123] is the LeNet-5 network. This architecture was developed specifically for digit recognition, marking a significant milestone in developing CNNs. The LeNet-5 architecture primarily consisted of alternating convolutional and pooling layers, culminating in a fully connected layer for classification. For many years, the basic structure of CNNs remained consistent until the advent of AlexNet [124], which significantly expanded both the depth and breadth of CNNs, introduced ReLU as activation functions, and popularized data augmentation as a technique for regularization. The success of AlexNet marked the beginning of the DL era, leading to numerous advancements in CNN architecture. VGGNet [125] and GoogLeNet [126] further advanced the field by building on the foundations established by AlexNet. VGGNet highlighted the importance of network depth for performance, while GoogLeNet introduced the inception module, which enabled the construction of deeper and wider networks without significantly increasing computational requirements. Figure 3.5 illustrates the basic architecture of the VGG-16 model.

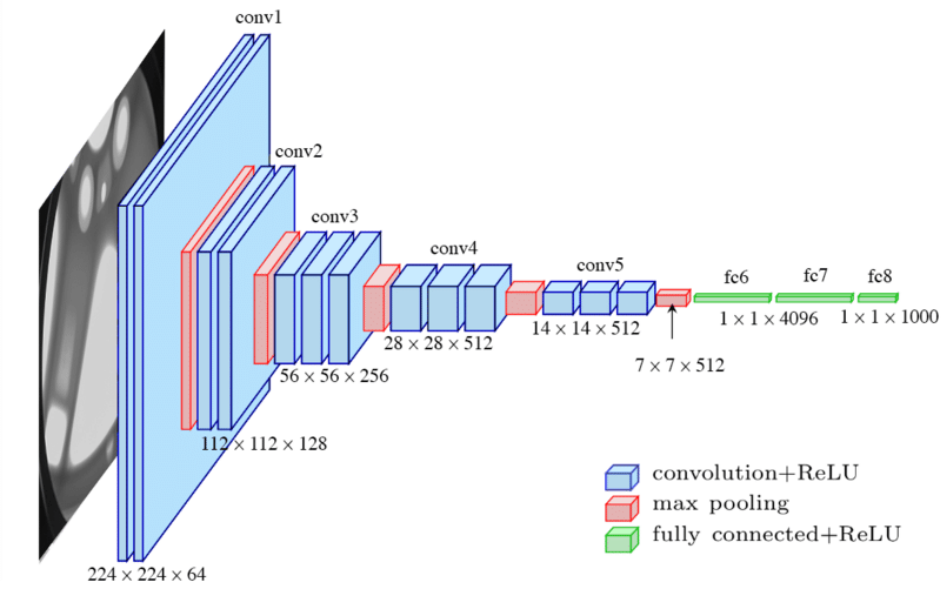

FIGURE 3.5: The standard VGGNet architecture as proposed in [125].

The release of ResNet [127] represented another significant milestone in developing CNNs and DL. ResNets introduced residual connections, which addressed the vanishing gradient problem and allowed for the successful training of deeper networks, even those with hundreds of layers. As CNNs continued to dominate in various computer vision tasks, researchers began to focus on making network architectures more efficient and scalable. This led to the development of MobileNet and EfficientNet [128], mainly designed to be both lightweight and high-performing. MobileNet introduced depthwise separable convolutions, while EfficientNet utilized a systematic scaling approach for depth, width, and resolution. More recently, RegNets [129] has employed a design space exploration approach to identify network architectures that offer optimal trade-offs between computational efficiency and accuracy. The latest advancement in efficient network design comes from NfNets [130], which have achieved state-of-the-art performance on ImageNet by employing a normalization-free design and model scaling, all while being more efficient than previous models.

Following the significant performance improvements achieved by CNNs in image classification [131], these models were adapted to tackle other computer vision tasks, such as semantic segmentation. Image segmentation involves assigning a specific label to each pixel in an input image. The primary objective is to detect, outline, and categorize every object within the image, ultimately generating a detailed prediction map. However, CNNs also faced challenges with these tasks. While pooling layers in deep models help expand the receptive field to gather contextual information and reduce the number of parameters, semantic segmentation requires that this context be maintained. In 2014, FCNs [132] became popular for making dense predictions without fully connected layers. They introduced skip connections between layers to combine coarse semantic context with local appearance details, thus enhancing the resolution during up-sampling. This end-to-end architecture enabled the generation of segmentation maps for images of varying sizes.

After the introduction of FCN, significant advancements were made in automatic segmentation techniques. These evolved architectures typically use a standard FCN to extract multi-scale features, followed by an upsampling stage that restores the input resolution using deconvolutional layers. The fully connected layers are omitted, and a pixel-wise classification layer is implemented at the end to create the final segmentation mask. This structure represents an early form of the convolutional *encoder-decoder* architecture, which has become the foundation for most modern semantic segmentation methods.

U-Net

U-Net, introduced by Ronneberger et al. [107], is a highly impactful architecture specifically designed for medical image segmentation, characterized by its symmetric encoder-decoder structure (Figure 3.6). The encoder path, also known as the contraction path, progressively reduces the spatial dimensions of the input image through a series of 3×3 convolutional layers, followed by ReLU activations and 2×2 max-pooling layers that downsample the data while doubling the number of feature channels at each step, allowing the network to capture increasingly complex features. The decoder path, or expansion path, mirrors this process, gradually restoring spatial dimensions through upsampling and convolutional operations. A key advancement in U-Net is skip connections, which directly connect corresponding layers in the encoder and decoder, allowing high-resolution features lost during downsampling to be effectively reintegrated into the segmentation map. These skip connections concatenate feature maps from the encoder with those in the decoder, ensuring that fine-grained detail is preserved, significantly improving segmentation accuracy, particularly in tasks requiring precise localization.

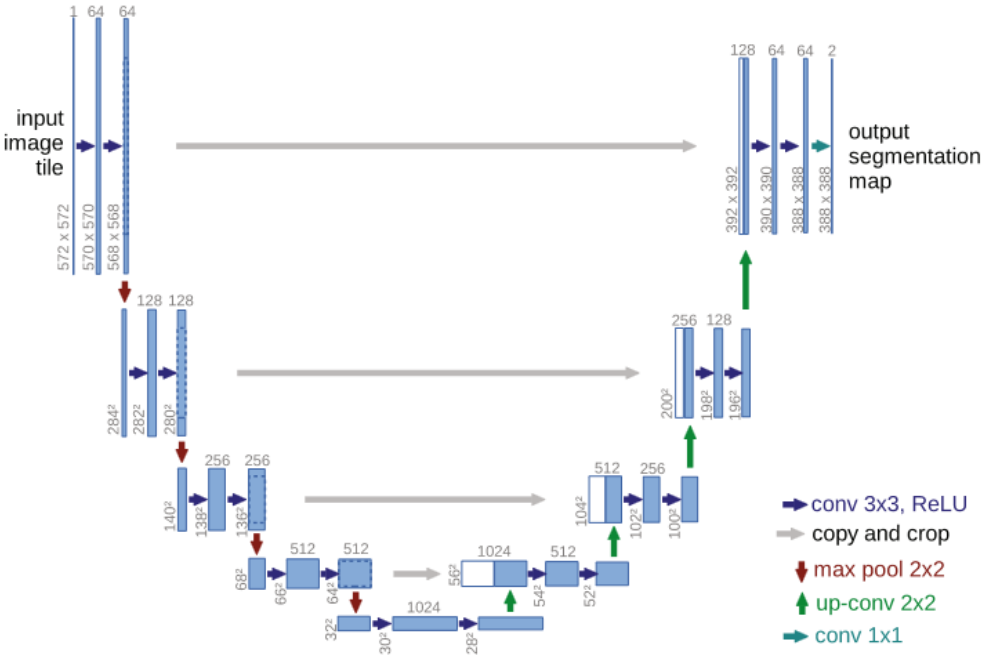

FIGURE 3.6: The standard U-Net architecture as proposed in [107].

U-Net Variants

Since the publication of U-Net in 2015, numerous variants have emerged, primarily aimed at enhancing its structural components while retaining its core symmetrical encoder-decoder structure. Advances include modifications such as FusionNet [133], which replaces conventional convolutional blocks with residual blocks [127] to improve feature extraction and gradient flow. CE-Net [134] introduces dense atrous convolutions and residual multi-kernel pooling to preserve high-level semantic information more effectively. V-Net [117], on the other hand, adapts the U-Net architecture for volumetric segmentation by utilizing 3D convolutions instead of 2D. Other variants add complexity to capture more details, such as M-Net [135], which integrates multi-scale inputs and deep supervision. UNet++ [136] builds upon the original U-Net by introducing nested skip pathways and dense skip connections between the encoder and decoder paths. These modifications enhance the network's ability to capture and integrate multi-scale features more effectively. UNet++ incorporates deep supervision by adding auxiliary output layers at various depths, which provides additional gradient signals and improves training convergence. Pohlen et al. [137] proposed an approach that incorporates a two-stream branch with pooling and residual streams. Fourure et al. [138] extend the standard unidirectional skip connections into a grid format, enabling feature maps of different shapes to interact more comprehensively. Additionally, in the vanilla U-Net, the information transmitted through skip connections is combined with the corresponding feature maps using addition. AttU-Net [139] is another prominent modification that integrates attention mechanisms into the standard U-Net framework. This variant introduces attention gates within the skip connections, enabling the network to focus on the most relevant features while suppressing less important information. By dynamically adjusting the weights of features based on their importance, AttU-Net improves the network's ability to capture fine details in complex images, making it particularly effective in scenarios where precise segmentation is crucial.

3.3.3 Vision Transformers

The development of the Transformer model to ViTs represents a significant evolution in DL architectures, extending the transformative impact of attention mechanisms from natural language processing (NLP) to computer vision. The Transformer model, introduced by Vaswani et al. [140], marked a significant advancement in NLP by using self-attention mechanisms to handle data sequences, effectively capturing dependencies and contextual information across long distances within the text. Unlike traditional sequence models that relied on recurrent or convolutional layers, the Transformer utilized

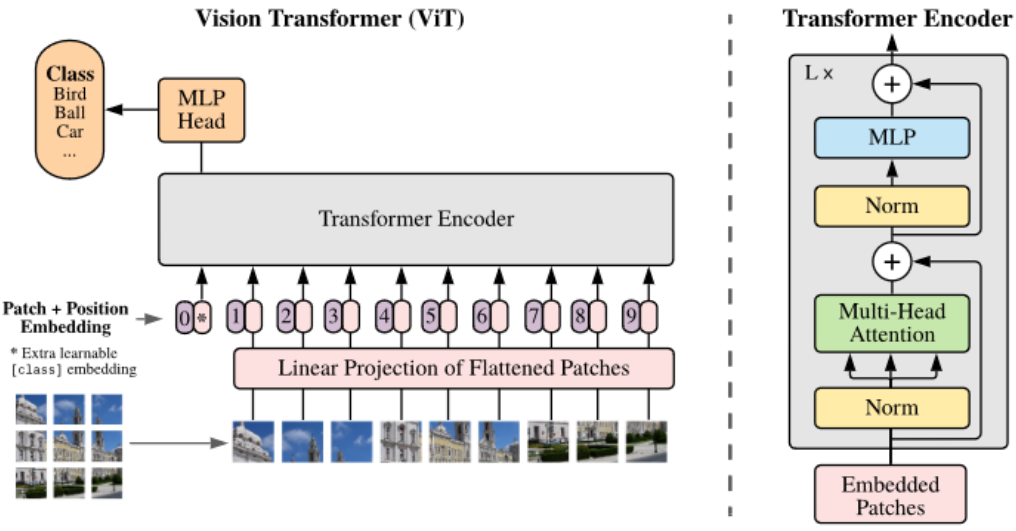

FIGURE 3.7: Architecture of the ViT as proposed in [142], with the detailed structure of the encoder block.

a purely attention-based approach, significantly improving training efficiency and model performance [141].

Building on this success, researchers explored the application of transformers to computer vision tasks, leading to the development of ViTs. Dosovitskiy et al. [142] adapted the transformer architecture for image processing by treating images as sequences of patches rather than pixels as illustrated in Figure 3.7. In this approach, an image is divided into fixed-size, non-overlapping patches, which are then linearly embedded into a sequence of tokens. These tokens are processed by the transformer layers, which consist of multi-head self-attention mechanisms and feed-forward networks, allowing the model to capture global dependencies and contextual information across the entire image. ViTs have achieved state-of-the-art results on multiple benchmarks [143], demonstrating the effectiveness of this switch from CNNs to transformer-based models for vision tasks. The main advantage of ViTs is their ability to model long-range relationships and contextual information across the entire image, which is often challenging for traditional CNNs due to their localized receptive fields [8].

3.3.3.1 Fundamental Components of ViT

The ViT model processes an image by dividing it into fixed-size patches, which are then flattened into vectors. These D-dimensional vectors are transformed by a trainable linear projection layer, resulting in N vectors, where N represents the number of patches. The resulting outputs, called patch embeddings, are combined with positional embeddings to preserve the spatial information of each patch. In addition, a trainable class embedding

is added to the patch embeddings before they are input into the Transformer encoder. The Transformer encoder consists of several blocks, each containing a multi-head self-attention (MSA) block and an MLP block. Before passing through these blocks, the activations are normalized using LayerNorm (LN). The model includes skip connections, which add a copy of the activations before LN to the outputs of the MSA or MLP blocks. Finally, an MLP block functions as a classification head, mapping the output to class predictions (Figure 3.7). One essential feature that sets Transformer models apart is the self-attention mechanism. Therefore, we begin by outlining the central idea of the attention mechanism.

Self-Attention

The attention mechanism was first introduced by Bahdanau et al. [144] for sequence-to-sequence models, particularly in neural machine translation. By dynamically focusing the model's "attention" on various parts of the input sequence based on their relevance to the task at hand, attention mechanisms improved the limitations of fixed-length context vectors and enhanced the learning of long-range dependencies.
In a self-attention layer (Figure 3.8 (a)), the input vector is initially transformed into three distinct vectors: the query vector q, the key vector k, and the value vector v, each with a fixed dimension. These vectors are combined into three separate weight matrices, denoted as W_Q, W_K, and W_V. For a given input X, the common forms of Q, K, and V can be expressed as:

$$K = W_K X, \quad Q = W_Q X, \quad V = W_V X \tag{3.14}$$

Where W_Q, W_K, and W_V are learnable parameters. The scaled dot-product attention mechanism is then defined by the following equation:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_K}}\right) V \tag{3.15}$$

Here $\sqrt{d_K}$ serves as a scaling factor, and the SoftMax function is applied to the attention weights to convert them into a normalized distribution.

Multi-Head Self-Attention

The MSA mechanism (Figure 3.8 (b)) was introduced to capture the complex relationships among token entities from various perspectives. Unlike single-head attention, which can be limited in capturing complex patterns, MSA enables the model to simultaneously focus on information from multiple representation sub-spaces. The MSA process

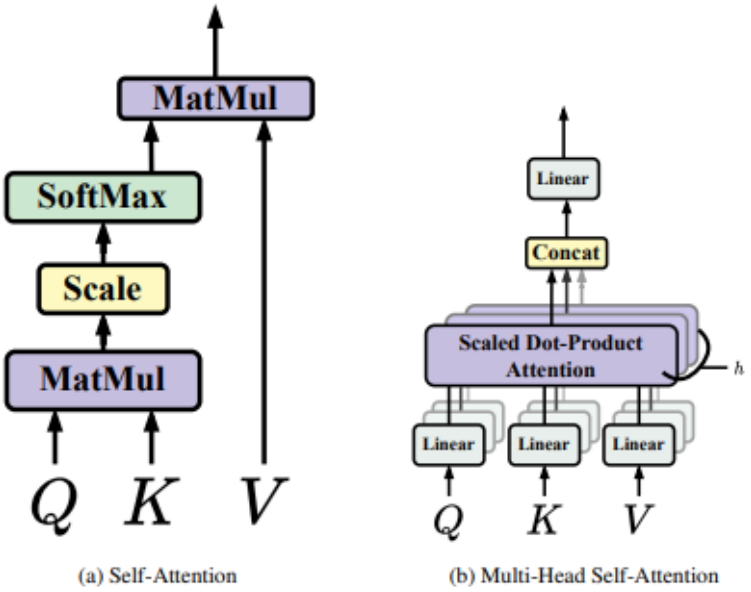

FIGURE 3.8: The structure of (a) Self-Attention and (b) Multi-Head Self-Attention as illustrated in [145].

can be represented as follows:

$$\text{MultiHead}(Q, K, V) = [\text{Concat}(\text{head}_1, \ldots, \text{head}_h)]W^O \tag{3.16}$$

Where each head is defined by:

$$\text{head}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i}) \tag{3.17}$$

Here W^O represents a linear mapping function used to combine the outputs of the multiple heads. The hyperparameter h is set to 8 in the original implementation.

3.3.3.2 ViT Architectures in Image Segmentation

While ViT was initially designed for image classification, its adaptability allows it to be extended to more complex tasks like image segmentation. Segmentation requires pixel-level predictions, signifying each pixel in the input image must be assigned to a specific class, necessitating modifications to the original ViT architecture. The most crucial adaptation involves replacing the classification head with a task-specific segmentation head capable of generating dense predictions for every pixel in the image. Unlike classification, where the output is a single class label, the segmentation head is designed to output a full-resolution segmentation map. This often involves implementing a decoder that upsamples the features from the Transformer encoder back to the original image resolution, ensuring that the model can accurately map its predictions to the precise

locations in the input image.
In addition to the segmentation head, positional encodings within the ViT play a critical role in maintaining the spatial relationships between image patches, which is essential for accurate segmentation. Some advanced segmentation models also incorporate a specialized decoder by employing transposed convolutions or a U-shaped architecture with skip connections to combine high-level abstract features with low-level detailed information [9]. These adjustments help ensure capturing the global context provided by the ViT's self-attention mechanisms and retaining the spatial detail necessary for precise segmentation [10]. Thus, while the core modification lies in adapting the classification head, additional architectural enhancements, like a dedicated decoder, can further refine the model's ability to handle the intricate demands of image segmentation tasks.
The Transformer architecture, initially introduced with an encoder-decoder structure, has since evolved significantly, resulting in numerous variants tailored to specific applications. ViTs are typically divided into two categories: pure and hybrid models [145]. A pure Transformer refers to using multiple multi-head self-attention modules in both the encoder and decoder, while a hybrid architecture combines ViTs with convolutional modules in various parts, such as the encoder, bottleneck, decoder, or skip connections, to integrate global context and local details.

Pure ViTs

These models, developed to address the limitations of CNN-based architectures in capturing global and long-range semantic information, are built entirely without convolutional layers. Instead, they rely on a structure comprising an encoder, bottleneck, decoder, and skip connections, all based directly on the ViT or its variants. These models feature multiple MSA modules within both the encoding and decoding stages, facilitating the decoder's utilization of the encoded information. Notable examples include the Swin-Unet [146] and SegFormer [147] networks.

Swin-UNet. Swin-Unet [146] is a medical image segmentation model with a symmetric encoder-decoder architecture based on the Swin Transformer block [148]. As shown in Figure 3.10, the Swin-Unet architecture is structured with an encoder, bottleneck, decoder, and skip connections. It uses the Swin Transformer block as its fundamental unit. In the encoder, medical images are divided into non-overlapping 4×4 patches having a feature dimension of 48 (i.e., $4 \times 4 \times 3$). These patches are then linearly embedded into a specific dimension, C, and passed through Swin Transformer blocks and patch merging layers to create hierarchical feature representations. The patch merging layers downsample the data and increase its dimensionality, while the Swin Transformer blocks handle feature representation learning. Each Swin Transformer block includes an LN

layer, an MSA, a residual connection, and a 2-layer MLP with GELU activation. The window-based multi-head self-attention (WMSA) module is used in one transformer block, while the shifted window-based multi-head self-attention (SWMSA) module is applied in the following block (Figure 3.9). Specifically, the operations within Swin Transformer blocks are represented as:

$$\hat{Z}^l = \text{WMSA}(\text{LN}(Z^{(l-1)})) + Z^{(l-1)} \tag{3.18}$$

$$Z^l = \text{MLP}(\text{LN}(\hat{Z}^l)) + \hat{Z}^l \tag{3.19}$$

$$\hat{Z}^{(l+1)} = \text{SWMSA}(\text{LN}(Z^l)) + Z^l \tag{3.20}$$

$$Z^{(l+1)} = \text{MLP}(\text{LN}(\hat{Z}^{(l+1)})) + \hat{Z}^{(l+1)} \tag{3.21}$$

Where $\hat{Z}^l$ and Z^l represent the outputs of the SWMSA/WMSA module and the MLP module of the l^{th} block, respectively.

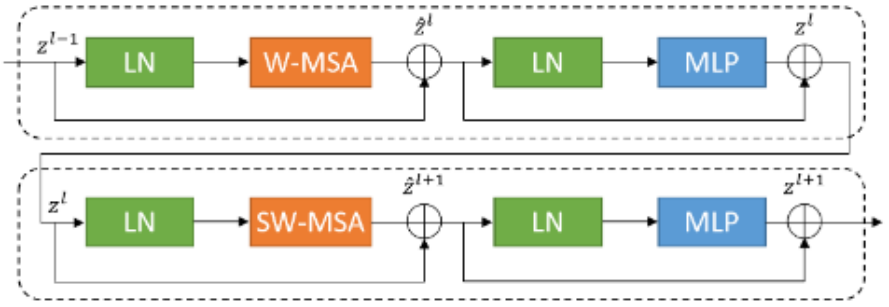

FIGURE 3.9: Swin Transformer block as illustrated in [146].

Inspired by U-Net, the architecture employs a symmetric transformer-based decoder composed of Swin Transformer blocks and patch-expanding layers. The decoder fuses context features with multiscale features from the encoder through skip connections, compensating for the spatial information loss during downsampling. Unlike patch merging layers, patch expanding layers perform upsampling by reshaping feature maps to double their resolution. Ultimately, the final patch expanding layer performs $4\times$ upsampling to restore the feature maps to the original input resolution ($W \times H$). Finally, a linear projection is applied to the upsampled feature maps to produce pixel-wise segmentation outputs. This configuration ensures that Swin-Unet effectively learns and integrates both local and global information, addressing challenges such as the loss of spatial detail that is common in pure Transformer models.

SegFormer. The SegFormer architecture [147], designed for semantic segmentation, consists of a hierarchical Transformer encoder and a lightweight All-MLP decoder. As illustrated in Figure 3.11, the process begins by dividing the input image into small

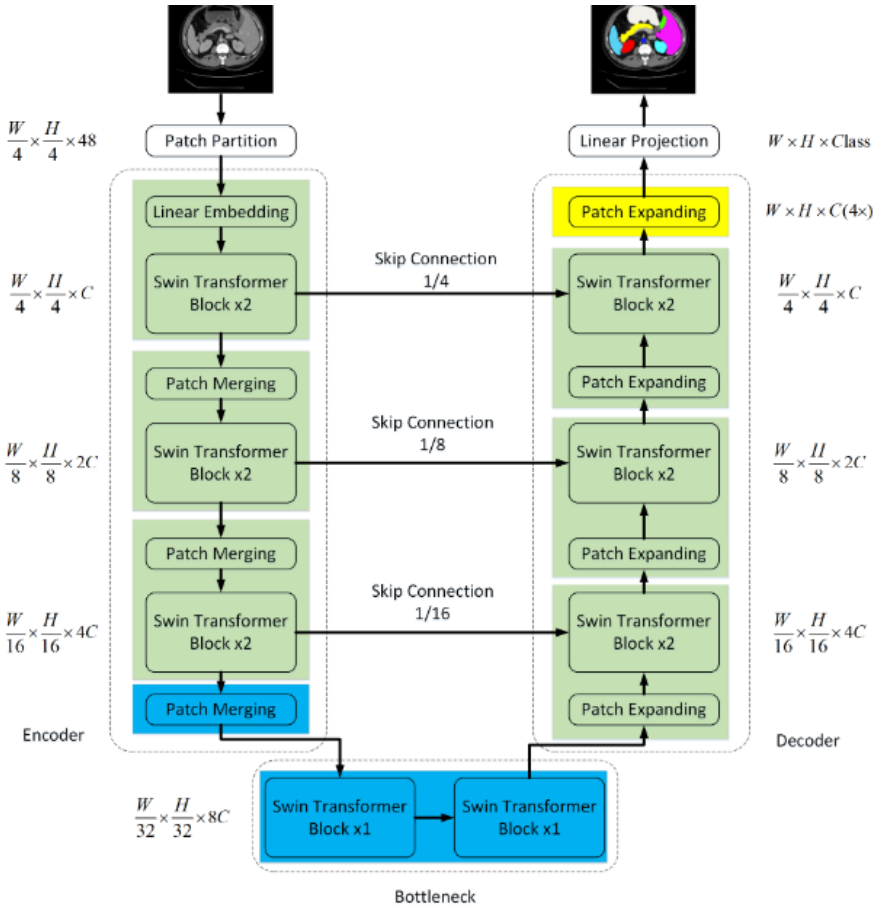

FIGURE 3.10: Overview of the SwinUNet architecture as proposed in [146].

patches of 4×4 pixels, fed into the hierarchical Transformer encoder. This encoder generates multi-level feature representations at various scales (1/4, 1/8, 1/16, and 1/32) of the original image size. The encoder, built on the Mix Transformer (MiT) design, uses overlapped patch merging to maintain local continuity and avoids using fixed positional encodings, which can be restrictive. Additionally, the self-attention mechanism is optimized through sequence reduction to manage the computational complexity associated with high-resolution features.

The decoder in SegFormer is composed entirely of MLP layers, making it lightweight and computationally efficient. It processes the multi-level features from the encoder by first unifying their channel dimensions, then upsampling and concatenating them. These features are further refined through MLP layers to produce the final segmentation mask at a resolution of $H/4 \times W/4 \times N_{\text{cls}}$, where N_{cls} is the number of segmentation categories. The architecture benefits from the large effective receptive field (ERF) produced by the Transformer encoder, which enables the decoder to capture local and global context information effectively without relying on complex, handcrafted components.

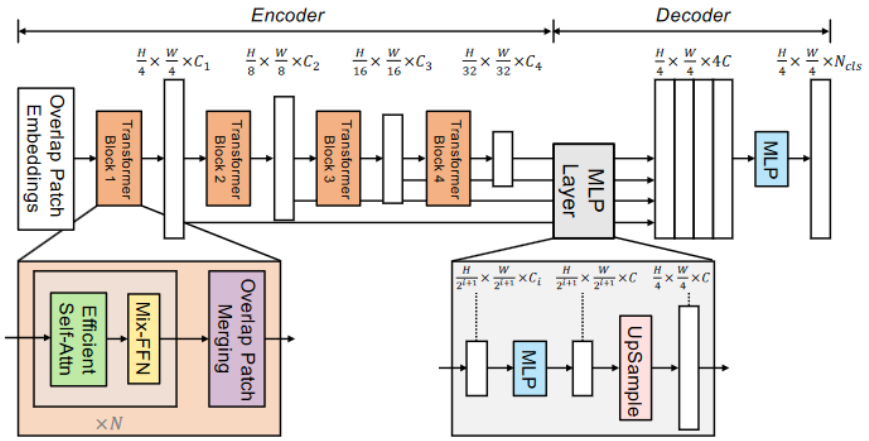

FIGURE 3.11: Overview of the SegFormer architecture as proposed in [147].

Hybrid ViTs

In segmentation tasks, hybrid models modify the traditional CNN structure by integrating Transformer modules at various stages within the architecture. These Hybrid Transformers combine Transformer blocks with convolutional layers to simultaneously capture local details and long-range dependencies. The integration can occur in different parts of the U-Net structure, such as the encoder, decoder, and skip connections. In the encoder, combining CNNs with Transformers enhances spatial detail recovery by addressing the challenge of insufficient low-level feature localization in pure Transformers [70, 149], [150, 151]. In the decoder, Transformers are used to refine the segmentation output, particularly in tasks involving generation [152]. The skip connections, which link the encoder and decoder, play a crucial role in enhancing network performance and convergence [153]. These connections ensure effective fusion and recalibration of features, ultimately leading to more precise and accurate segmentation outcomes.

TransUNet. TransUNet [70] is an advanced architecture designed to leverage the strengths of both CNNs and Transformers in medical image segmentation tasks. In this architecture, CNNs are placed at the beginning of the encoder to extract and encode local features from the input images (Figure 3.12). CNNs excel at capturing fine-grained spatial details due to their localized receptive fields, making them ideal for producing feature maps that highlight important structures within the image. These feature maps are then fed into the Transformer block within the same encoder. The Transformer is specifically employed to capture long-range dependencies and global context across the image. This is achieved through its self-attention mechanism, which allows the model to consider the entire image context when refining the encoded features. By placing the Transformer block after the CNN, TransUNet ensures that the local details are preserved while also benefiting from the broader context that Transformers provide.

The decoder of TransUNet further enhances this combination by using a cascaded upsampling process, which incrementally reconstructs the high-resolution segmentation map. Skip connections are introduced between the encoder and decoder to maintain a seamless flow of information, enabling the network to fuse global context from the Transformer with the fine details initially captured by the CNN. This organization allows TransUNet to achieve superior segmentation performance by balancing detailed feature localization with global contextual understanding, making it particularly effective for complex medical imaging tasks.

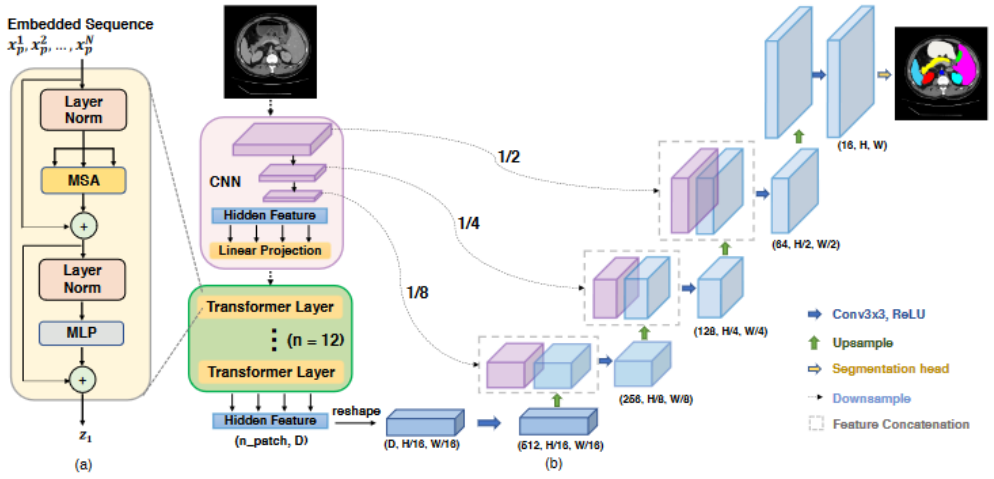

FIGURE 3.12: Overview of the TransUNet architecture as proposed in [70].

3.3.4 Optimization of Deep Learning Models

DL models, such as CNNs and ViTs, have demonstrated significant potential in imaging-based CAD systems. However, their direct application in clinical settings faces several challenges. These models are typically resource-intensive, requiring substantial computational power and memory, which can limit their deployment on standard clinical hardware or edge devices. Additionally, the trade-off between model complexity and performance complicates their use: more complex models may provide higher accuracy but often result in slower response times and increased costs, making them less suitable for time-sensitive medical environments.

To overcome these challenges, TL techniques offer a promising solution. TL can improve the performance and generalization of DL models by transferring knowledge from pre-trained models, enabling faster training and better accuracy without requiring as much computational power or data. This approach helps make DL models more efficient and effective, enhancing their viability in clinical applications.

3.3.4.1 Transfer Learning

The rapid evolution of machine learning has led to numerous new applications. While supervised learning achieves strong performance due to large datasets, some domains struggle with limited or inaccessible data, making it difficult to train effective models. This challenge has given rise to TL, which involves pretraining models on extensive datasets like ImageNet [154] and then fine-tuning them for specific tasks. TL has been widely adopted across various fields, including image recognition [155], language processing, and medical applications like tuberculosis detection, pneumonia diagnosis via chest X-rays, breast cancer classification, and cardiac image segmentation [156]. Key techniques in TL encompass instance-based, network-based [157], homogeneous, heterogeneous [158], and adversarial approaches [156]. Among them, instance-based and network-based TL approaches are often selected to solve practical application problems [159].

Domains and tasks are fundamental concepts in TL. In this context, a domain D comprises two components: a feature space X and a marginal probability distribution $P(X)$, which together define the domain as:

$$D = \{X, P(X)\} \tag{3.22}$$

A task T in TL is defined by a label space Y and a target prediction function $f(X)$, which is also interpreted as the conditional probability $P(Y \mid X)$. This is represented as:

$$T = \{Y, P(Y \mid X)\} \tag{3.23}$$

In TL, the source domain refers to the domain used to train the model or perform tasks, while the target domain is where the trained model is applied for predicting, classifying, and clustering data.

Instance-Based TL

Instance-based TL leverages instances from a source domain D_s by applying a tailored weighting or selection mechanism. This approach involves identifying and incorporating certain instances from the source domain D_s into the target domain D_t training process, effectively enriching the target data set. A key strategy within this approach is the TrAdaBoost method [160], which systematically filters out instances from D_s that do not closely align with the characteristics of D_t, ensuring that only relevant and similar instances are used. This method enhances the model's ability to generalize in D_t by focusing on the most pertinent instances from D_s. The instance-based transfer can be

represented as:

$$f\big((D_s \to T_s) \times (D_t \to T_t)\big) \to f\big((D_s \to D_t) \to T_t\big) \tag{3.24}$$

Network-Based TL

Network-based deep TL involves leveraging pretrained neural networks to address tasks in a target domain D_t. In this approach, the first layers of the neural network, originally trained on a source domain D_s, are repurposed as robust feature extractors. These layers, with their established structures and parameters, are integrated into a new network tailored for D_t. Often, these layers remain unchanged or frozen during the retraining process, ensuring that the extracted features from D_s provide a strong foundation for the target task. This method allows for efficient knowledge transfer, reducing the need for extensive data in the target domain while maintaining high performance [161]. The network-based transfer reuses model weights and parameters in T_s as:

$$f\big((D_s \to T_s) \times (D_t \to T_t)\big) \to f\big(D_t \to (T_s \to T_t)\big) \tag{3.25}$$

3.4 Conclusion

This chapter provided an in-depth examination of the general process of image-based CAD systems, outlining key stages such as preprocessing, assisted diagnosis (including classification and segmentation), and decision-making. It explored DL techniques, emphasizing the basic components and architectures of CNNs and ViTs, along with their applications in medical image segmentation tasks. In addition, optimization strategies, including TL techniques, were discussed, highlighting their role in improving the efficiency, scalability, and performance of DL models in real-world medical imaging scenarios.

Chapter 4

Intelligent Mask Image Reconstruction for Cardiac Image Segmentation through Local-Global Fusion

4.1 Introduction

Cardiac image segmentation is crucial for diagnosing and treating heart disease, providing detailed insights into the structure and function of the heart. Accurate segmentation of cardiac cine MRI images allows for a thorough analysis of the RV, LV, and Myo, which is essential for evaluating cardiac function and anatomy. However, the manual segmentation process is labor-intensive and prone to errors due to fatigue, making it a time-consuming task.

DL has greatly advanced the automation of image segmentation, particularly in cardiac imaging, where techniques like ViTs have significantly improved performance [64, 162, 163]. ViTs excel at capturing complex patterns within images, leading to better accuracy in tasks such as cardiac cine MRI segmentation [164–167]. Despite their advantages, Transformers can struggle with capturing fine details [145], which has led to the development of various strategies to enhance their performance. These include hybrid architectures combining Transformer and Convolutional blocks [70, 149], improving pure Transformer architectures [146, 150], and integrating Transformers with other DL approaches like CNNs [168, 169]. Despite significant progress and the development of many robust models, no single segmentation model consistently outperforms others across all scenarios. Models that achieve high accuracy on one dataset often struggle to

deliver similar results on other datasets or when the same dataset is expanded. Furthermore, the potential of integrating different types of Transformers within a single model to effectively address these challenges remains to be explored in current research.
This chapter introduces three significant contributions to address these challenges:

- A novel ensemble framework, FCTransNet, is introduced for accurate segmentation of cardiac structures in cine MRI, integrating the capabilities of three state-of-the-art ViT models.
- We present an advanced pixel-level image fusion approach, called Intelligent Weighted Summation Technique (IWST), which constructs a final segmentation mask by merging the output of each ViT model. This fusion strategy exploits the complementary strengths and different perspectives of each model.
- Extensive evaluations on the ACDC dataset show that FCTransNet outperforms current ViT-based methods and other DL approaches, delivering superior segmentation results in a competitive setting.

The rest of this chapter is structured as follows: Section 4.2 offers an overview of recent related works on using ViTs in cardiac image segmentation and medical image fusion techniques. Section 4.3 outlines the methods employed. Section 4.4 presents and examines the experimental results. Section 4.5 provides an in-depth discussion of the findings, and Section 4.6 concludes with future directions and perspectives.

4.2 Related Works

Transformers have become a powerful tool for segmenting cardiac images, successfully overcoming challenges that traditional DL methods often encounter. These models excel in capturing long-range dependencies, maintaining critical spatial details, incorporating global context, and adapting well to situations with limited data. Previous research has mainly focused on three strategies: first, enhancing pure Transformer models by modifying or replacing certain blocks; second, creating hybrid architectures by merging Transformers with CNNs; and third, creating ensemble learning approaches combining Transformers with other DL models like CNNs [145].
One of the key approaches' researchers have explored is enhancing pure ViTs architectures for cardiac image segmentation by introducing or replacing specific blocks within the transformer structure to better suit the unique characteristics of cardiac images. For instance, Cao et al. [146] introduced Swin-Unet, the first U-shaped architecture solely based on transformers, which utilizes Swin transformer blocks in the encoder,

decoder, and bottleneck stages, incorporating patch merging and expanding layers for down-sampling and up-sampling while retaining the original skip connection for feature processing. Huang et al. [170] developed the MISSFormer model, which modifies the transformer structure by introducing Efficient Self-Attention for high-resolution features. Zhou et al. [171] presented nnFormer, a model based on a modified transformer for 3D medical images that utilizes local and global transformer blocks in the encoder, decoder, and bottleneck stages to learn 3D volumes at both local and global scales. Liu et al. [172] introduced the TransFusion model, which incorporates two new modules within a transformer model to address complex medical image segmentation tasks by constructing semantic dependencies across diverse scales and views. Galazis et al. [173] proposed Tempera, a feature pyramid with a geometric spatial transformer for multiple passes, employing a spatial transformer to establish connections between various views of the heart, facilitating smooth transitions between 2D and 3D images. Recently, Gao et al. [174] proposed Medformer, a transformer designed for scalable 3D medical image segmentation with broad generalizability. Unlike existing vision transformers, MedFormer integrates three essential components: a favorable inductive bias, hierarchical modeling featuring linear-complexity attention, and multi-scale feature fusion to integrate semantic and spatial information. This model effectively learns from limited medical data and generalizes across various medical imaging tasks without pre-training.

Further, the integration of Transformer blocks with CNN architectures has been explored to improve cardiac image segmentation. A notable example is the incorporation of transformer blocks into U-Net architectures, which have been highly effective in biomedical image segmentation. The first work to integrate transformer blocks into the U-Net structure for cardiac image segmentation was presented by Chen et al. [70] in their TransUNet model. This model features an encoder-decoder structure, where the encoder uses a transformer-based self-attention mechanism to capture the overall context, while the decoder employs U-Net-like skip connections for precise segmentation. Building on TransUNet, Xu et al. [75] introduced LeViT-UNet, the first medical image segmentation model to feature transformer blocks, by integrating the LeViT [175] as a new encoder. Additionally, Yang et al. [176] developed TransNUNet, a new transformer-based segmentation model derived from TransUNet, with an encoder structure identical to that of TransUNet. Gao et al. proposed two transformer-based models, UTNet [177] and UTNetV2 [174], to address the high computational costs associated with medical image segmentation. UTNet applied transformer modules at every layer (except the first), incorporating an efficient MSA module to reduce complexity while preserving image details through transformer decoders. UTNetV2 further improved efficiency by integrating a bidirectional transformer block in both the encoder and decoder, facilitating the fusion of multi-scale feature information. These models demonstrated superior performance in capturing long-range dependencies, accurately segmenting boundaries, and integrating

multi-scale semantic maps. Deng et al. [178] introduced TransBridge, a network architecture that minimizes computational costs by shortening the token sequence length and replacing skip connections with transformer encoders on specific layers. This approach enhances the integration of feature maps from CNN layers, achieving comparable segmentation performance with significantly fewer parameters than existing models. Wu et al. [179] introduced D-Former for 3D image segmentation, which incorporates local and global scope modules to reduce computational complexity, with LS-MSA and GS-MSA processing local units and using dilated methods for connections between units, alongside 3D depth-wise convolution to further lower computational costs. Additionally, Transformers are integrated into the model to leverage their strong attention mechanisms for capturing long-range dependencies and incorporating global context. The combination of these models (i.e., Transformers and U-shaped models) has been shown to improve accuracy and segmentation results compared to traditional CNN-based approaches.

In addition to integrating Transformers and CNN blocks, researchers have also explored the parallel combination of transformers with CNNs. For example, Luo et al. [180] proposed a semi-supervised framework that combines CNNs and Transformers through a cross-teaching strategy for medical image segmentation. This method addresses the challenge of achieving good performance with limited annotations by using the predictions of one network as labels to supervise the other network in an end-to-end manner. Wang et al. [166] utilized two Vision Transformers for cardiac image segmentation, where the proposed approach comprises student and teacher models. The student model assists the teacher in parameter updates by learning from image features, aiming to minimize both supervision loss for segmentation and semi-supervision loss for consistency by ensuring the agreement between the inference of unlabeled data by the two models.

In the past, most research in medical image segmentation, especially cardiac image segmentation, has focused on combining transformers with existing architectures, enhancing standalone transformer models, and integrating transformers with other deep learning techniques like CNNs. However, our work takes a different approach by uniting various transformer-based models, exploring ensemble techniques, and utilizing the unique strengths of different transformer architectures to effectively tackle challenges in cardiac segmentation.

4.3 FCTransNet

This section provides a comprehensive explanation of the proposed FCTransNet approach, as illustrated in Figure 4.1. The system consists of two key modules: an ROI extraction module utilizing a custom U-Net architecture and a fusion module that merges the outputs of several ViTs through the proposed image fusion technique.

FCTransNet is a multi-transformer framework designed to segment anatomical structures in cardiac cine MRI scans by integrating three powerful ViT models. The process begins with a 2D preprocessed cine MRI image fed into a custom U-Net model responsible for localizing and extracting the ROI containing critical structures such as the LV, Myo, and RV. Next, the extracted ROI is processed by three ViTs (TransUNet, SwinUNet, and SegFormer) each tasked with segmenting specific cardiac structures. The final segmentation output is achieved by combining the predicted masks from these models using a new pixel-level medical image fusion method called the Intelligent Weighted Summation Technique (IWST). This technique intelligently integrates information from the three predicted masks, accounting for both the global features and the local context of each pixel, ensuring a thorough and adaptive fusion process that refines the final segmentation in FCTransNet.
To our knowledge, this is the first study to combine different Vision Transformers for image segmentation and the first to apply pixel-level image fusion to merge masks from different segmentation models.

4.3.1 ROI Extraction Module

The initial step of FCTransNet involves isolating a specific ROI that includes crucial cardiac structures such as the LV, Myo, and RV (Figure 4.1). This extraction serves multiple objectives. Firstly, it confines the analysis to relevant areas by removing unnecessary background regions, which could otherwise introduce noise and impede the training effectiveness of the segmentation models. Secondly, this phase reduces the computational complexity of FCTransNet by eliminating the need to process images in their original, larger dimensions. Additionally, the ROI extraction addresses the prevalent issue of imbalanced classes in medical image segmentation [181]. By excluding excessive background regions, the prominence of the foreground structures is enhanced, thereby alleviating the imbalance problem.
In this phase, an enhanced U-Net architecture is employed to detect and extract the ROI. The model generates masks that guide the selection of the optimal bounding box for the targeted structures. Before being fed into the U-Net, the 2D slices undergo preprocessing, which includes two main steps: resizing and normalization. The slices are first resized to 224×224 pixels. Slice-wise normalization is then applied to standardize the pixel intensity values across the slices. This normalization adjusts the pixel values to fall within a range of 0 to 1, using the following formula:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4.1}$$

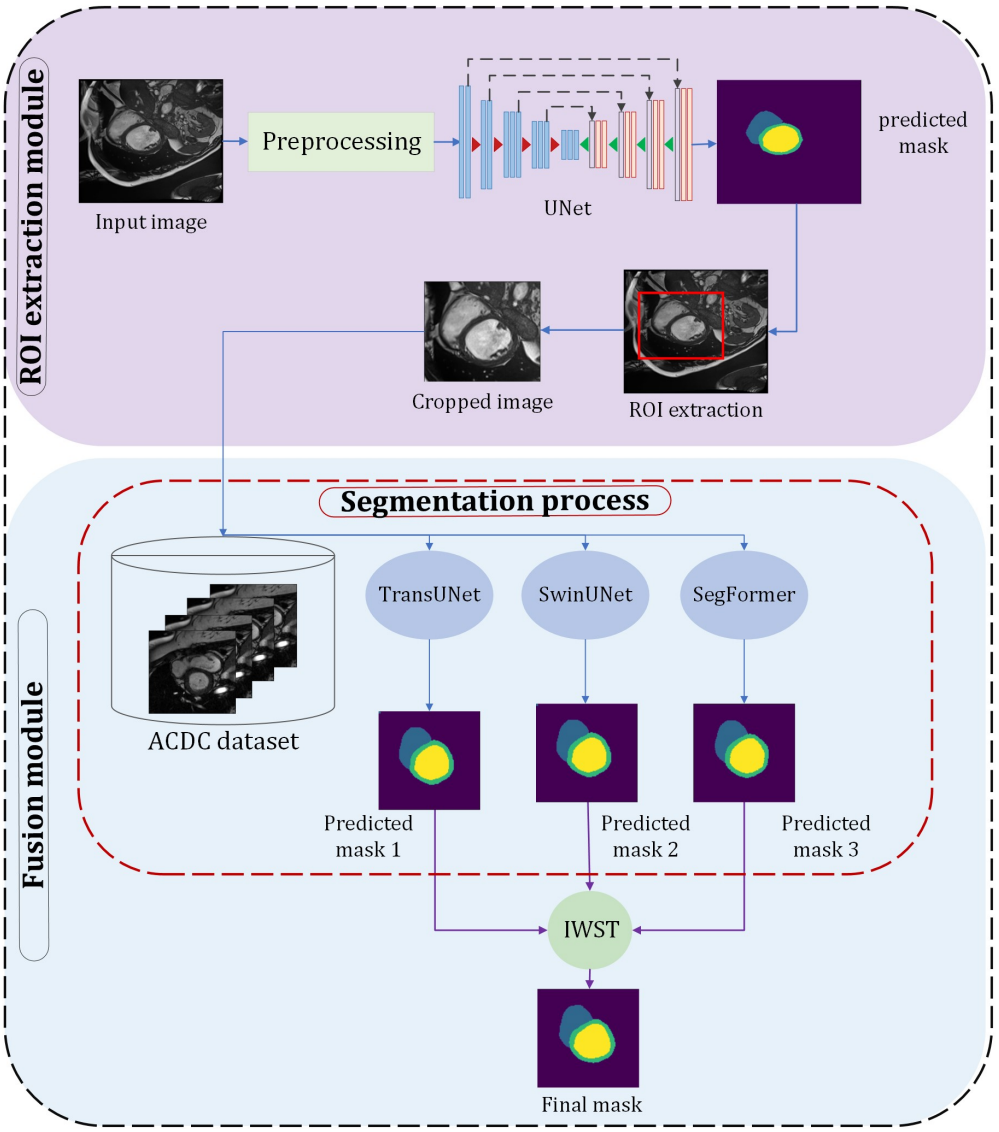

FIGURE 4.1: Overview of the FCTransNet framework.

Here, x represents the pixel value in the current slice, x_{norm} is the normalized pixel value, and $x_{\min}$ and $x_{\max}$ denote the minimum and maximum pixel intensities in that slice. This procedure ensures consistency in pixel intensity across the slices, which is crucial for the following stages of the approach.

The U-Net model used here follows a simple encoder-decoder architecture with long skip connections (as shown in Figure 4.2). The encoder consists of four downsampling blocks inspired by the VGG-16 model, which help in capturing the context. Each block includes Convolutional layers followed by Max Pooling and ReLU activation. The network incorporates 13 convolutional layers based on VGG-16. In the decoder, each upsampling block includes Upsampling, Convolutional, and Batch Normalization (BN) layers with ReLU activation. BN layers are applied after each convolutional layer in the decoder, followed by ReLU activation. Long skip connections combine the detailed features from

the encoder with the broader features in the decoder, aiding in the recovery of spatial details lost during downsampling. These skip connections are vital for producing accurate segmentation masks and improving the network's training process. The final layer uses a 1×1 convolution followed by a sigmoid activation function to classify the pixels, with the output ranging from 0 to 1, where 1 indicates the structure of interest.

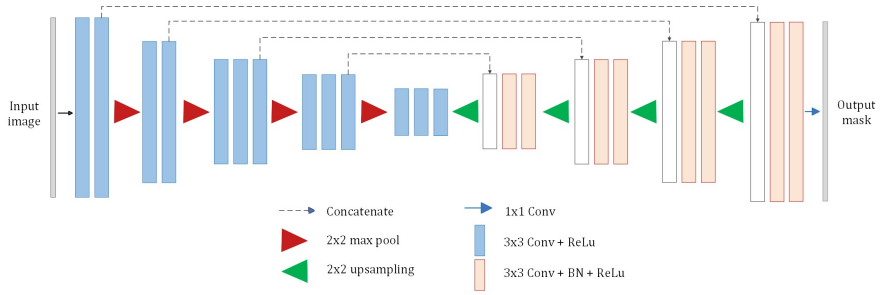

FIGURE 4.2: The enhanced UNet architecture for the extraction of the ROI from cardiac images.

4.3.2 Base Models of FCTransNet

The core ViT architectures that serve as the foundation of our proposed approach form the basis of FCTransNet. These ViTs will be combined through a fusion process in the following stage of our approach. We have selected and adapted three renowned and highly effective ViTs (TransUNet [70], SwinUNet [146], and SegFormer [58]) to serve as the foundational models for FCTransNet.

In Section 2.2, we explored various transformer-based models in medical image processing. These models can be classified into three principal categories: pure transformer-based methods, hybrid architectures that combine CNNs with transformer blocks, and ensemble approaches that combine CNNs and transformer models. Our proposed method adopts a unique combination of ViTs to construct a new framework. The chosen ViTs represent a diverse range of architectures, including two pure transformer models and one hybrid model that incorporates both CNNs and transformer blocks. This selection is driven by two main considerations:

1. **Effectiveness in Medical Image Segmentation**: The selected ViTs have demonstrated strong performance in addressing the challenges of medical image segmentation [150, 182, 183]. Their ability to accurately identify and differentiate anatomical structures within medical images has been consistently validated [145]. These transformers excel in extracting relevant features from complex medical data, contributing to reliable and clinically meaningful segmentation results.

2. **Architectural Variety**: The distinct architectural designs of these transformers highlight their unique strengths and limitations in solving similar problems. Combining these three robust segmentation transformers aims to benefit from their complementarity to create a more comprehensive and effective solution for the segmentation of cardiac images.

4.3.3 IWST-Based Fusion Module

This module's primary function is to combine the predictions generated by the individual base transformers, specifically TransUNet, SwinUNet, and SegFormer.
Subsequently, the outputs from these base transformers are integrated into the fusion module. Within this specialized module, the three predicted images, each originating from a different base transformer, are integrated. Drawing inspiration from existing pixel-level image fusion methods, we devised a novel approach known as the IWST to facilitate this integration. Figure 4.3 provides an overview of the fusion module.

4.3.3.1 Intelligent Weighted Summation Technique

The Intelligent Weighted Summation Technique (IWST) is an advanced pixel-level image fusion method designed to integrate information from n input images. The primary objective of IWST is to generate a final fused image, denoted as F, that effectively combines pixel information from the input images using a smart weighting mechanism. The key innovation of IWST, in contrast to traditional weighted summation methods, lies in its consideration of a local window surrounding each pixel rather than focusing only on the individual pixel itself. Given two input images, X and Y, the IWST method calculates the fusion of two pixels at a specific location (i, j) according to the following formula:

$$F(i,j) = \frac{1}{(2m+1)^2} \sum_{k_2=-m}^{m} \sum_{k_1=-m}^{m} (W_X \times X(i+k_1, j+k_2) + W_Y \times Y(i+k_1, j+k_2)) \tag{4.2}$$

Where $2m+1$ represents the size of the local window, m can be 0, 1, or 2. $F(i,j)$ is the fused pixel value at the coordinates (i,j) in the resulting image. W_X and W_Y are weights assigned to the corresponding pixels from images X and Y, respectively. These weights determine the influence of each pixel value in the fusion process.
The IWST technique allows for the targeted highlighting of particular features or details by assigning tailored weights to various regions within the images, facilitating the integration of complementary information from both sources. Weights W_X and W_Y are

constructed from two partial weights, W_g and W_l.

$$W_X = W_g \times W_l \tag{4.3}$$

$$W_Y = W_g \times W_l \tag{4.4}$$

The weight W_g represents the global view provided by a particular model's mask, offering a comprehensive overview of the entire image and capturing its global features. This weight is crucial in highlighting or downplaying specific regions based on the characteristics identified by the model, ensuring that the fusion process considers the overall context of the input images. On the other hand, W_l represents the class to which a pixel belongs, providing a localized view. This allows the fusion process to account for the specific class information of each pixel, enabling the technique to adjust its weighting according to the significance of different classes during the fusion. The multiplication of W_g and W_l ensures that both global and local information contributes to the pixel weighting process during the fusion stage. This combined weighting method enables a more sophisticated integration of features from the model's mask and the class-specific details of each pixel, thereby enhancing the technique's adaptability and capacity to capture contextual nuances in image fusion.

To clarify our proposed method, let's consider an example: Imagine two pixels, x and y, located at position $(20, 20)$ in images X and Y. We set the parameter m to 1, defining a local window with a size of $2m + 1$, which equals 3×3. As illustrated in Figure 4.3, the IWST technique applies this window, where blue and yellow squares correspond to the selected regions from images X and Y. The fused pixel value, f, is calculated by averaging the values within the green square, representing the sum of the two windows. This example demonstrates how IWST combines information from neighboring pixels in both images to produce a fused pixel value at the given location, enabling a weighted integration of surrounding features.

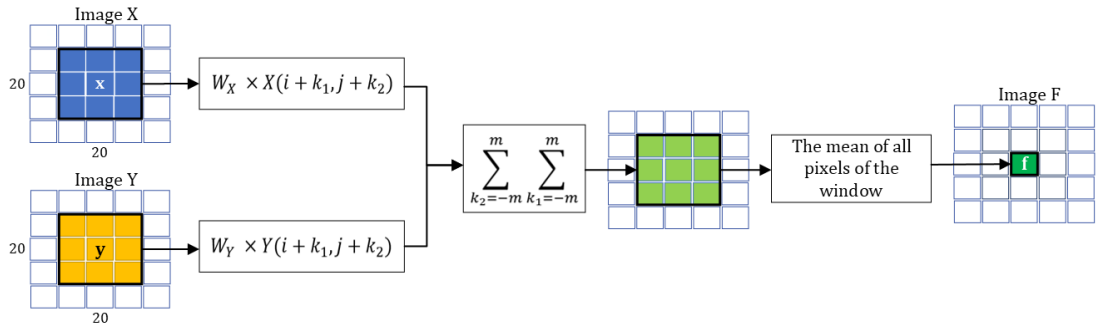

FIGURE 4.3: Illustration of IWST technique application: Example with window size (2m+1)=3, and pixel position (20,20).

The IWST algorithm (Algorithm 1) computes the fused pixel value by applying a weighted sum to the corresponding pixels within the local window of the input images. The weights W_X and W_Y enable the algorithm to dynamically adjust the significance of each input image's contribution, ensuring that crucial information from both images is retained throughout the fusion process.

In essence, the IWST method employs a combination of local windowing and weighted summation to effectively blend information from two input images, producing a fused image that embodies the complementary features of the original images.

Algorithm 1 Intelligent Weighted Summation Technique (IWST)

Require: Images X and Y, weights W_X and W_Y, local window size $2m + 1$ (where $m \geq 0$), pixel coordinates (i, j)
Ensure: Fused pixel value $F(i, j)$
Initialize $\text{sum}_F \leftarrow 0$
for $k_2 \leftarrow -m$ to m **do**
    for $k_1 \leftarrow -m$ to m **do**
        $\text{sum}_F \leftarrow \text{sum}_F + W_X \times X(i + k_1, j + k_2) + W_Y \times Y(i + k_1, j + k_2)$
    end for
end for
$F(i, j) \leftarrow \frac{1}{(2m+1)^2} \times \text{sum}_F$
return $F(i, j)$

4.4 Experimental Results

4.4.1 Dataset

The Automated Cardiac Disease Diagnosis Challenge (ACDC) dataset [58], introduced during the 2017 MICCAI conference, is a significant resource in the medical imaging community, serving as the first and largest publicly available fully annotated cardiac MRI dataset. The dataset includes images from 150 patients, carefully categorized into five distinct subgroups based on various heart diseases: Normal, Myocardial Infarction, Dilated Cardiomyopathy, Hypertrophic Cardiomyopathy, and Abnormal Right Ventricle. Each subgroup contains 30 patients, ensuring balanced representation across different cardiac conditions. Each scan represents 3D volumes created from 2D short-axis cine images, with an in-plane pixel size varying between 1.34 to 1.68 mm^2, a slice thickness ranging from 5 to 10 mm, and occasional 5 mm interslice gaps. The volumes typically contain 6 to 18 slices, capturing between 28 to 40 phases per cardiac cycle. These images were acquired using a 1.5T or a 3.0T MRI scanner, specifically the Siemens Aera or Siemens Trio Tim models from Siemens Medical Solutions, Germany, covering the heart from base to apex.

For each patient, a single expert manually segmented the 2D cine slices at both the ES and ED phases, targeting the RV, Myo, and LV. Two independent experts reviewed and double-checked these segmentations, reaching a consensus on the final annotations to ensure accuracy.

After extracting 2D slices, we obtained 1,489 2D images from 150 3D volumes, excluding those with blank ground truth (Figure 4.4). The resulting dataset is split into training, testing, and validation sets in a 70:20:10 ratio. The problem of segmenting cardiac structures from short-axis cine MRI was approached as a four-class segmentation task using a 3-channel input.

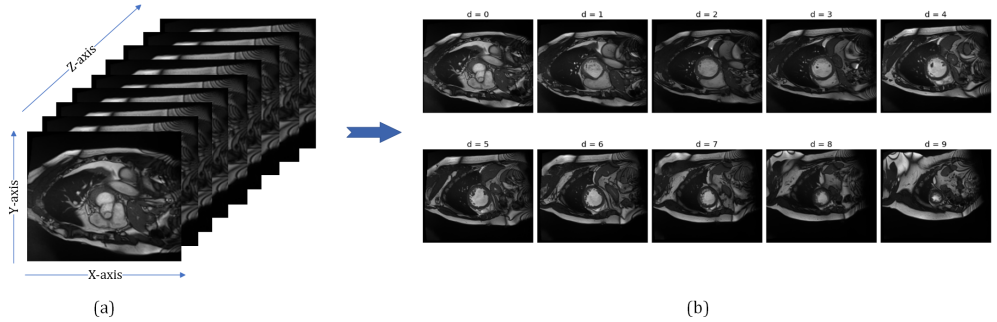

FIGURE 4.4: Process of Extracting 2D Slices from a 3D Cardiac MRI Volume. (a) 3D cardiac MRI volume representation with slices along the z-axis. (b) Example of 2D slices extracted from the 3D volume at different depth levels (d=0 to d=9), illustrating anatomical changes as the slices move through the z-axis.

4.4.2 Implementation Details

The performance of FCTransNet was evaluated compared to the individual ViTs, TransUNet, SwinUNet, and SegFormer as well as the robust UNet segmentation network [108, 184]. Each model was reimplemented according to the specific details in their respective papers. All experiments were conducted using Python 3.10.10, TensorFlow 2.11.0, and Windows 10. Training was performed on an NVIDIA Tesla P100 GPU with 16 GB of RAM. Unlike previous approaches that used pre-trained models on ImageNet, the base models in FCTransNet were initialized randomly.

The DL networks were trained on pairs of 2D MR images and ground truth labels. Before being inputted into the ROI extraction module, which employs a UNet architecture, the 2D slices underwent preprocessing consisting of two main steps: resizing and normalization. The slices were first resized to 224×224 pixels, followed by slice-wise normalization (Eq. 4.1) to ensure consistency across image slices. The resulting ROIs were then reduced to 128×128 pixels. To balance the dataset, increase the variety of training examples, enhance the model's generalization ability, and reduce the risk of overfitting, augmentations like horizontal flips, rotations, and grid distortions were independently

applied to both images and their corresponding ground truth. The patch size was set to 4, with a consistent batch size of 4 and 100 epochs for all models to ensure stable convergence. The Adam optimizer was used for network optimization, with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}.
In medical image segmentation, the class imbalance between the region of interest and the surrounding background is a common challenge that can hinder accurate model training. To address this, a hybrid loss function combining cross-entropy loss and Dice loss was used for training. Cross-entropy loss measures the pixel-wise error between predicted and target classes, while Dice loss evaluates the overlap between the predicted and ground truth segmentation. By integrating these two loss functions, the model can balance overall classification accuracy with segmentation precision. The total loss function is weighted with a ratio of 2:3 for cross-entropy and Dice loss, respectively:

$$L_{\text{Total}} = 2 \cdot L_{\text{CE}} + 3 \cdot L_{\text{Dice}} \tag{4.5}$$

4.4.3 Experimental Results and Ablation Studies

4.4.3.1 Comparative Analysis with Related Works

This section presents an objective evaluation of FCTransNet's performance on the ACDC Challenge dataset, followed by a comparison with leading state-of-the-art methods. We divide the comparative analysis into two categories: (a) studies that employed transformer-based models for cardiac structure segmentation (Table 4.1) and (b) studies that utilized alternative deep learning techniques (Table 4.2).

Quantitative Evaluation

To quantitatively assess the segmentation performance against transformer-based approaches, we utilized the DSC, the 95th percentile Hausdorff Distance (HD95), and the IoU to evaluate the accuracy of cardiac structure segmentation during the ED phase. For studies employing other deep learning methods, the DSC and HD were applied to both the ED and ES phases. The results obtained by FCTransNet on the ACDC test dataset, along with comparisons to other relevant studies, are shown in Tables 4.1 and 4.2.
The proposed method demonstrated strong performance, particularly in segmenting the LV, Myo, and RV. Among transformer-based approaches, FCTransNet achieved superior outcomes across all metrics, surpassing advanced models such as nnFormer [171] and D-Former [179]. For example, FCTransNet excelled in DSC, HD95, and IoU. Notably, the method showed significant improvements in segmenting LV and RV, especially

in challenging cases with complex shapes and intensity variations within the ventricles (Figure 4.6 (a)). According to Table 4.1, FCTransNet achieved an average DSC of 0.985, reflecting a 6.4% increase over nnFormer and a 6.2% improvement over D-Former (Figure 4.6 (b)). The average HD95 was 1.000 mm, 0.120 mm lower than that of nnFormer, while the average IoU was 0.914, surpassing the second-best by 12.5%. Figure 4.6 illustrates the comparison of DSC values between FCTransNet and other transformer-based methods on the ACDC test dataset, highlighting a significant advantage of FCTransNet over the second-best approach, D-Former. The difference in DSC is represented by the red and blue lines in Figure 4.6.

Our approach ranks among the top three in achieving high DSC values for LV, Myo, and RV segmentation during the ED phase and surpasses all existing methods by achieving the best DSC and HD values during the ES phase. Specifically, as shown in Table 4.2, FCTransNet achieved the third-highest DSC of 0.964 for LV and the second-highest DSC of 0.961 for RV in the ED phase, along with the lowest HD values of 2.828 for LV and 3.000 for Myo. In the ES phase, FCTransNet led with the highest DSC values of 0.950 for LV, 0.933 for Myo, and 0.930 for RV, as well as the best HD scores of 2.236 for LV, 3.000 for Myo, and 7.280 for RV. Comparative charts for DSC values related to RV, Myo, and LV segmentation during the ED and ES phases are provided in Figure 4.7.

Qualitative Evaluation

Figure 4.5 presents a visual comparison of consecutive cardiac MR slices, showcasing the input image, ground truth, UNet, TransUNet, SwinUNet, SegFormer, and our proposed FCTransNet. In this figure, the red, blue, and green labels correspond to RV, LV, and Myo, respectively. The ground truth and the segmentation results produced by our method are displayed in the second and last columns of Figure 4.5. It is noteworthy that FCTransNet is able to accurately segment the target regions across all five slices, whereas other models are unable to consistently achieve this within the same input image. Yellow rectangles are used to highlight these failure areas, which manifest as either missing regions within the target areas or as unintended segmentation in the background. This visual analysis is consistent with the earlier quantitative findings.

Methods	**Year**	**Average**			**LV**		**Myo**		**RV**	
		DSC	**HD95**	**IoU**	**DSC**	**HD95**	**DSC**	**HD95**	**DSC**	**HD95**
TransUNet [70]	2021	0.897	-	-	0.957	-	0.845	-	0.889	-
SwinUNet [146]	2021	0.900	-	-	0.958	-	0.856	-	0.886	-
MISSFormer [185]	2021	0.908	-	-	0.950	-	0.880	-	0.896	-
LeVit-UNet [75]	2021	0.903	-	-	0.938	-	0.876	-	0.896	-
ECT-NAS [186]	2021	0.898	-	-	0.888	-	0.850	-	<u>0.957</u>	-
Luo et al. [180]	2021	0.911	3.600	-	0.941	4.000	<u>0.893</u>	2.400	0.900	4.400
D-Former [179]	2022	<u>0.923</u>	-	-	<u>0.959</u>	-	**0.896**	-	0.913	-
UTNet V2 [174]	2022	0.917	3.550	-	0.946	2.910	**0.896**	2.490	0.910	5.270
nnFormer [171]	2022	0.921	<u>1.120</u>	-	0.957	<u>1.090</u>	**0.896**	<u>1.040</u>	0.909	<u>1.230</u>
Wang et al. [166]	2022	0.882	-	<u>0.789</u>	-	-	-	-	-	-
ATFormer [187]	2023	0.874	2.090	0.778	0.906	1.540	0.850	1.720	0.850	1.720
MCRformer [188]	2023	0.908	-	-	0.954	-	0.886	-	0.885	-
FCTransNet (Ours)	2024	**0.985**	**1.000**	**0.914**	**0.964**	**1.000**	0.892	**1.000**	**0.961**	**1.000**

TABLE 4.1: Segmentation accuracy of FCTransNet in comparison with transformer-based methods and frameworks. The highest and second-highest performances are highlighted in bold and underlined text, respectively.

Methods	ED						ES					
	LV		Myo		RV		LV		Myo		RV	
	DSC	**HD**	**DSC**	**HD**	**DSC**	**HD**	**DSC**	**HD**	**DSC**	**HD**	**DSC**	**HD**
Isensee et al. (2018) [189]	0.967	5.476	**0.904**	7.014	0.946	**8.205**	0.928	6.921	0.919	7.328	0.904	11.655
Baumgartner et al. (2018) [190]	0.963	6.526	0.892	8.703	0.932	12.670	0.911	9.170	0.901	10.637	0.883	14.691
Zotti et al. (2019) [191]	0.964	6.180	0.886	9.586	0.934	11.052	0.912	8.386	0.902	9.291	0.885	12.650
Painchaud et al. (2019) [192]	0.961	6.152	0.881	8.651	0.933	13.718	0.911	8.278	0.897	9.598	0.884	13.323
Khened et al. (2019) [68]	0.964	8.129	0.889	9.841	0.935	13.994	0.917	8.968	0.898	12.582	0.879	13.930
Simantiris & Tziritas (2020) [193]	0.967	6.366	0.891	8.264	0.936	13.289	0.928	7.573	0.904	9.575	0.889	14.367
Baldeon Calisto & Lai-Yuen (2020) [163]	0.958	5.592	0.873	8.197	0.936	10.183	0.903	8.644	0.895	8.318	0.884	12.234
da Silva et al. (2022) [194]	0.963	8.062	0.894	7.906	0.900	14.660	0.912	10.432	0.905	9.912	0.860	17.560
Dong et al. (2022) [195]	**0.970**	7.000	**0.904**	9.000	0.949	12.200	0.935	8.300	0.919	11.500	0.898	12.600
Wang et al. (2022) [196]	0.944	-	0.896	-	**0.965**	-	0.892	-	0.912	-	0.926	-
FCTransNet (Ours)	0.964	**2.828**	0.892	**3.000**	0.961	15.095	**0.950**	**2.236**	**0.933**	**3.000**	**0.930**	**7.280**

TABLE 4.2: Segmentation accuracy of FCTransNet in comparison with other DL-based approaches. Bold and underlined text denote the best and second-best performances, respectively.

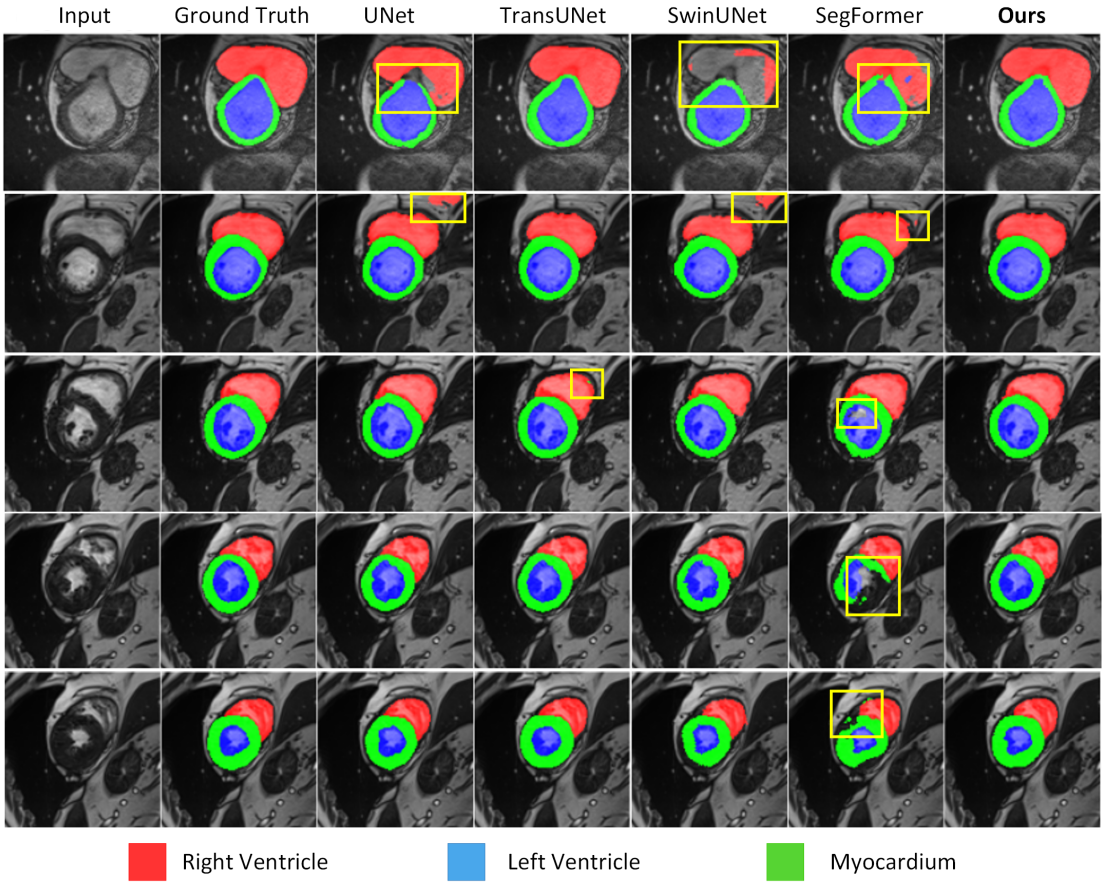

FIGURE 4.5: Visualization of segmentation results produced by different methods using ED phase test data from the ACDC dataset. The regions representing RV, Myo, and LV are highlighted in red, green, and blue, respectively.

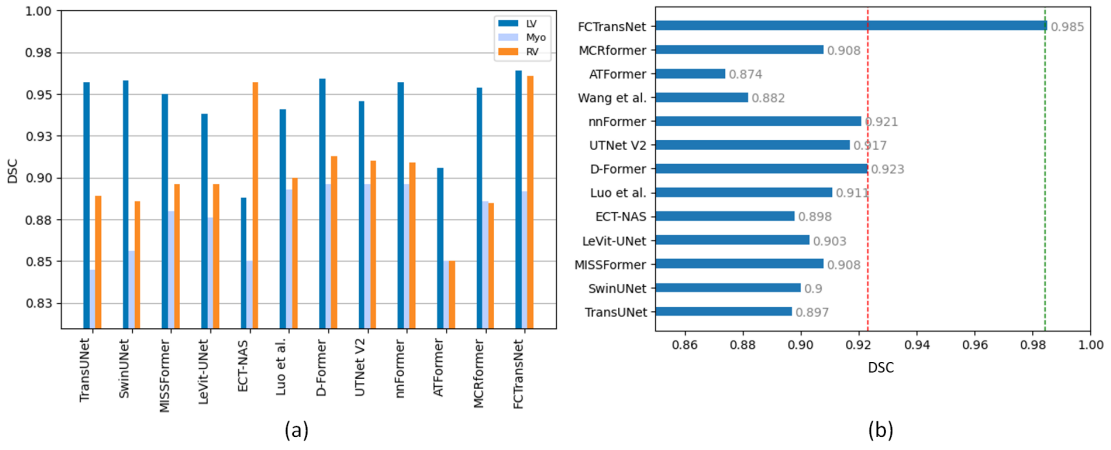

FIGURE 4.6: Bar chart comparing the DSC for transformer-based methods: (a) DSC values for each structure (RV, Myo, and LV), and (b) the average DSC. The red and blue lines in (b) indicate the difference between the average DSC of our proposed approach, FCTransNet, and the second-highest average DSC (D-Former).

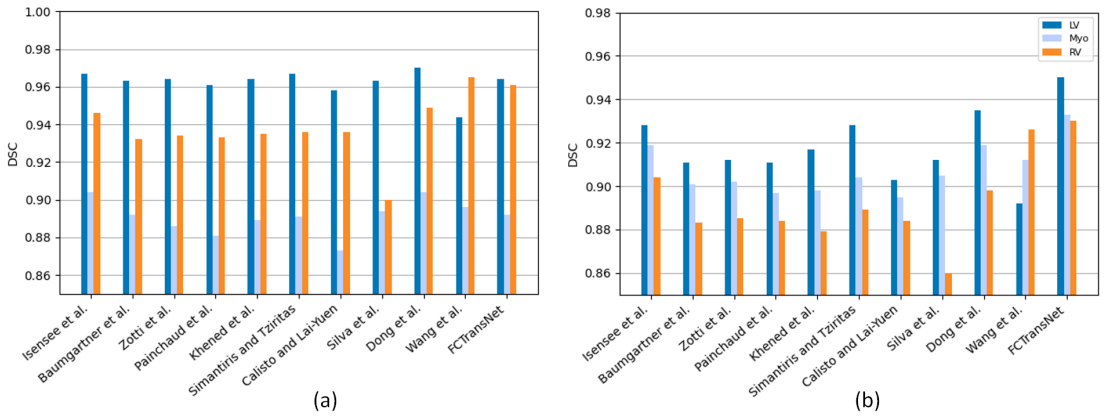

FIGURE 4.7: Bar chart comparing DSC values for various deep learning-based methods: (a) DSC for RV, Myo, and LV structures during the ED phase, and (b) DSC during the ES phase.

4.4.3.2 Analysis Study

In this section, we present an in-depth analysis of the significance and potential benefits of our proposed image fusion technique, IWST, in comparison to three existing methods: averaging, weighted averaging, and voting. Tables 4.3 and 4.4 showcase the segmentation performance of the IWST technique against the other fusion methods during the ED and ES phases, respectively. The results indicate that IWST outperforms the other methods, achieving a mean DSC of 0.958, HD of 9.854, ASSD of 0.161, and IoU of 0.914 during the ED phase, and a mean DSC of 0.989, HD of 4.172, ASSD of 0.165, and IoU of 0.908 during the ES phase.

The qualitative segmentation results on the ACDC test dataset are presented in Figure 4.8. It can be clearly seen that IWST technique outperforms other image fusion methods. Although the averaging, weighted averaging, and voting methods demonstrate good segmentation performance on the ACDC test dataset, they exhibit inaccuracies, particularly in the RV structure in the first and second images. In contrast, the IWST technique achieves perfect performance across all five images.

As illustrated in Figure 4.9, a line graph visually displays the experimental results, with the x-axis indicating the names of the image fusion techniques and the y-axis showing the metric values. The IWST model achieves optimal segmentation performance based on various experimental indicators shown in the line chart.

Fusion method	Average				LV			Myo			RV		
	DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
Averaging	0.982	12.835	0.172	0.904	0.961	5.000	0.146	0.885	5.000	0.171	0.956	21.095	0.138
Weighted averaging	0.979	11.835	**0.161**	0.911	0.963	3.000	0.132	0.890	3.000	0.160	0.958	21.095	**0.132**
Voting	0.975	11.872	0.171	0.904	0.960	3.605	0.130	0.891	**2.828**	0.159	0.942	20.808	0.182
IWST technique	**0.985**	**9.854**	**0.161**	**0.914**	**0.964**	**2.828**	**0.129**	**0.892**	3.000	**0.154**	**0.961**	**15.095**	**0.132**

TABLE 4.3: Comparison of FCTransNet segmentation performance during the ED phase using the IWST technique versus other fusion methods.

Fusion method	Average				LV			Myo			RV		
	DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
Averaging	0.984	4.195	0.171	0.902	**0.951**	2.239	0.127	0.931	**3.000**	0.151	0.921	7.348	0.240
Weighted averaging	0.986	4.175	0.175	0.903	0.950	2.238	0.125	0.930	**3.000**	0.147	0.924	7.285	0.229
Voting	0.982	4.595	0.170	0.901	0.950	2.236	0.129	0.928	3.162	0.160	0.930	7.385	0.230
IWST technique	**0.989**	**4.172**	**0.165**	**0.908**	0.950	**2.236**	**0.122**	**0.933**	**3.000**	**0.147**	**0.930**	**7.280**	**0.226**

TABLE 4.4: Comparison of FCTransNet segmentation performance during the ES phase using the IWST technique versus other fusion methods.

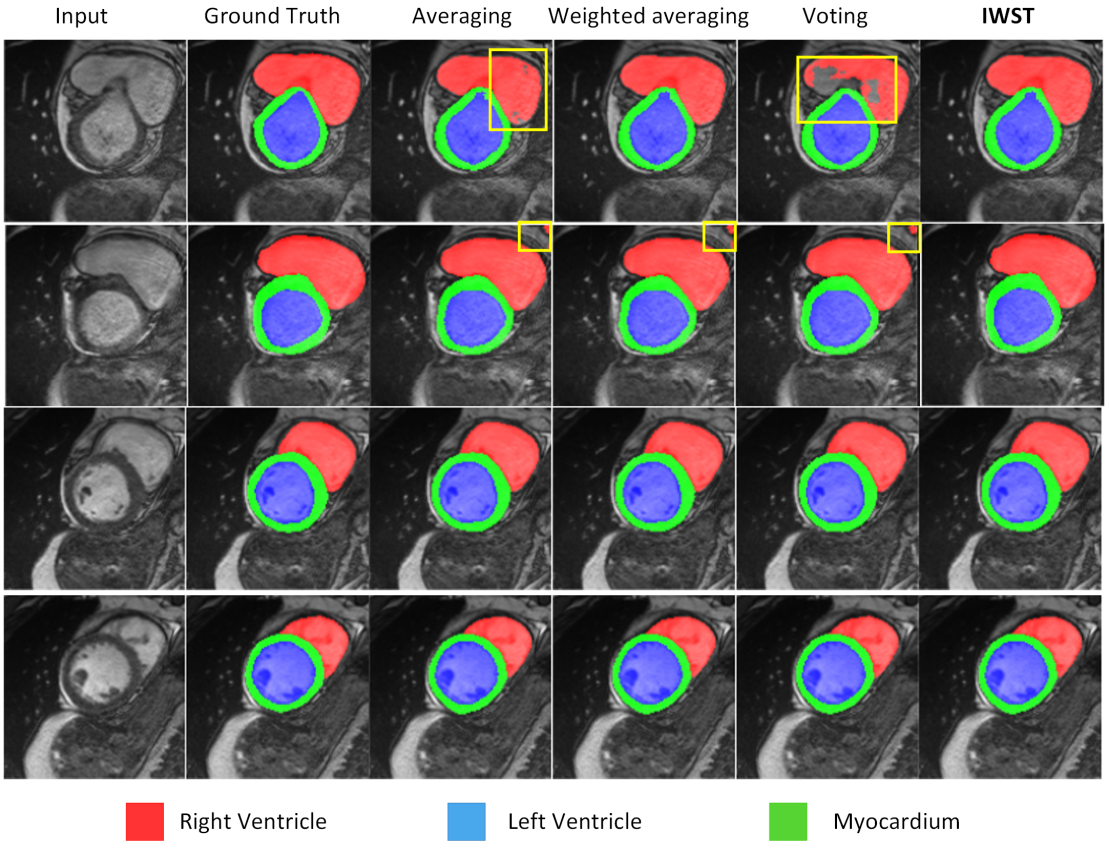

FIGURE 4.8: Line chart comparing the IWST technique with other image fusion methods.

4.4.3.3 Ablation Experiments

This section presents a series of experiments designed to evaluate the performance of the ROI extraction module, a critical component of the proposed framework. Through meticulous ablation studies, we analyze how incorporating the ROI extraction module and the proposed UNet architecture influences segmentation accuracy, highlighting their contributions to the overall efficacy of our method while identifying potential areas for improvement.

Tables 4.5 and4.6 provide a quantitative comparison of segmentation outcomes with (W) and without (W/o) the ROI extraction module for the UNet, three base transformers, and FCTransNet during both the ED and ES phases. These tables highlight the substantial impact of the ROI extraction module on the performance of individual transformers and the overall method. Specifically, the integration of the ROI extraction module resulted in notable improvements in segmentation performance for FCTransNet, with increases in DSC, HD, ASSD, and IoU of 0.006, 1.041, 0.262, and 0.111, respectively, in the ED phase, and 0.007, 10.68, 1.659, and 0.365, respectively, in the ES phase.

FIGURE 4.9: Visualization of segmentation results for FCTransNet with various image fusion techniques including averaging, weighted averaging, voting, and IWST on the ACDC test dataset during the ED phase. The regions corresponding to RV, Myo, and LV are presented in red, green, and blue, respectively.

Model		**Average**				**LV**			**Myo**			**RV**		
		DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
UNet	W/o	0.958	14.910	0.514	0.751	0.893	11.358	0.386	0.674	10.770	0.506	0.824	21.601	0.651
	W	0.972	12.343	0.287	0.885	0.957	7.141	0.164	0.860	3.605	0.212	0.940	19.313	0.311
TransUNet	W/o	0.974	20.036	0.740	0.808	0.906	16.125	0.918	0.749	18.191	0.311	0.871	19.496	0.896
	W	0.983	15.120	0.163	0.911	0.963	3.000	**0.125**	**0.895**	16.309	0.160	0.958	16.095	0.145
SwinUNet	W/o	0.954	15.095	0.866	0.609	0.768	13.454	0.850	0.512	20.050	0.740	0.657	17.247	0.959
	W	0.967	14.476	0.370	0.849	0.950	5.000	0.199	0.845	18.000	0.229	0.880	16.633	0.449
SegFormer	W/o	0.961	21.180	1.463	0.752	0.897	21.827	1.091	0.709	21.475	1.115	0.794	21.652	0.940
	W	0.972	11.835	0.229	0.872	0.941	15.588	0.214	0.838	8.485	0.237	0.946	**5.000**	0.142
FCTransNet	W/o	0.979	10.895	0.423	0.803	0.921	3.000	0.219	0.778	7.211	0.295	0.875	24.269	0.539
	W	**0.985**	**9.854**	**0.161**	**0.914**	**0.964**	**2.828**	0.129	0.892	**3.000**	**0.154**	**0.961**	15.095	**0.132**

TABLE 4.5: Segmentation results from the ablation study of the ROI extraction module during the ED phase. "W/o" indicates results without the module, while "W" denotes results with the module.

Model		**Average**				**LV**			**Myo**			**RV**		
		DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
UNet	W/o	0.979	11.696	0.861	0.681	0.818	12.884	0.479	0.801	11.225	0.508	0.548	16.454	1.504
	W	0.981	7.580	0.199	0.889	0.938	4.472	0.167	0.915	3.742	0.197	0.913	15.033	0.236
TransUNet	W/o	0.979	17.652	1.704	0.537	0.660	17.378	1.208	0.655	12.728	0.691	0.516	20.421	3.328
	W	0.987	16.158	0.234	0.907	0.947	2.236	0.137	0.931	3.000	0.147	0.924	19.698	0.311
SwinUNet	W/o	0.975	15.652	1.704	0.537	0.660	17.378	1.208	0.655	12.728	0.691	0.516	20.421	3.328
	W	0.979	12.389	0.525	0.861	0.934	4.123	0.166	0.888	3.606	0.243	0.880	10.915	0.798
SegFormer	W/o	0.960	15.107	4.330	0.413	0.433	18.619	5.599	0.412	21.254	1.859	0.548	21.699	4.376
	W	0.974	7.779	0.315	0.852	0.931	3.606	0.180	0.892	4.472	0.261	0.855	13.038	0.464
FCTransNet	W/o	0.982	14.852	1.824	0.543	0.691	15.297	1.006	0.642	12.042	0.677	0.618	20.712	3.960
	W	**0.989**	**4.172**	**0.165**	**0.908**	**0.950**	**2.236**	**0.122**	**0.933**	**3.000**	**0.147**	**0.930**	**7.280**	**0.226**

TABLE 4.6: Segmentation results from the ablation study of the ROI extraction module during the ES phase. "W/o" indicates results without the module, while "W" denotes results with the module.

Figure 4.10 provides a visual illustration of segmentation results that highlights the impact of excluding the ROI extraction module from FCTransNet. The absence of this critical component leads to a noticeable drop in segmentation quality, affecting both precision and accuracy. Consequently, the performance of the model is less effective compared to when the full FCTransNet, with the ROI extraction module included, is used. These observations underscore the importance of the ROI extraction module in improving the model's segmentation performance.

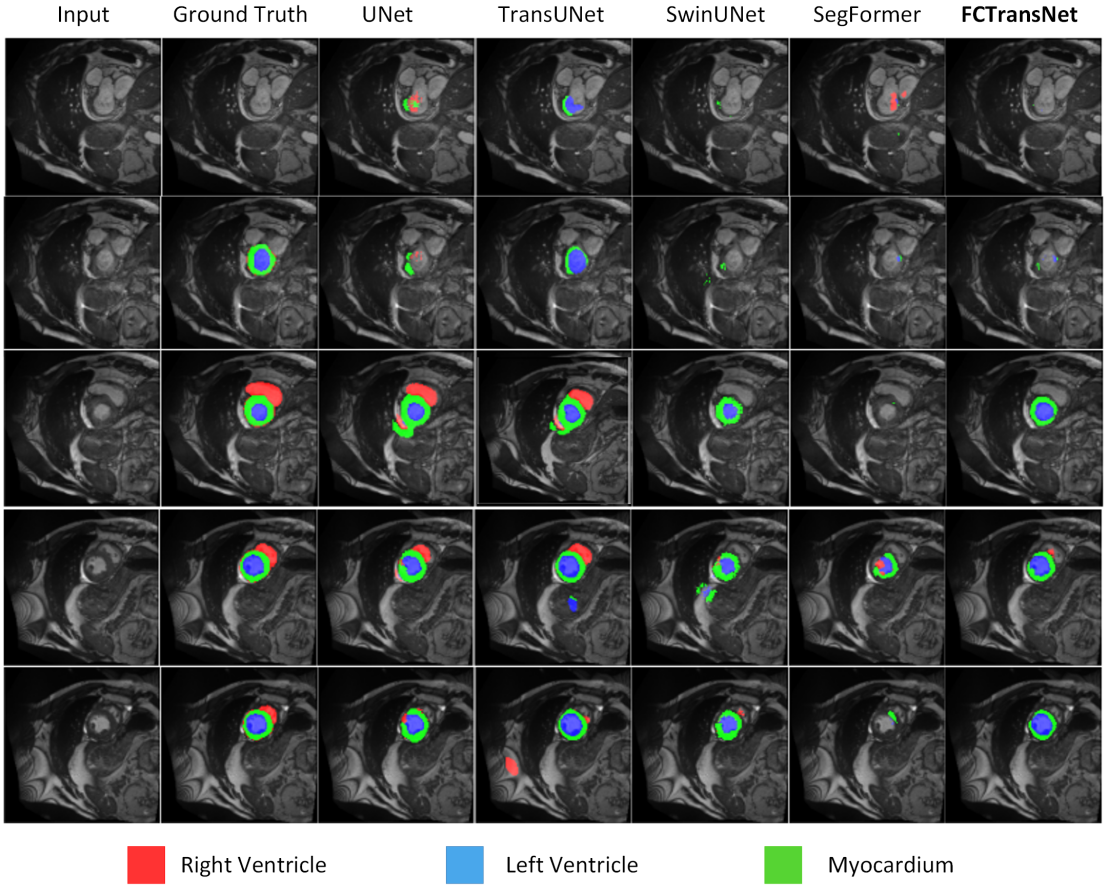

FIGURE 4.10: Visualization of segmentation results from various individual models and the proposed FCTransNet on the ACDC test dataset during the ES phase, without ROI extraction. The regions for RV, Myo, and LV are indicated in red, green, and blue, respectively.

A detailed examination of Figure 4.11 illustrates a discernible pattern highlighting the effectiveness of the ROI extraction module. This module is crucial in substantially boosting the performance of the FCTransNet method in both the ED and ES phases. This emphasizes the module's important role in enhancing the method's performance during these stages.

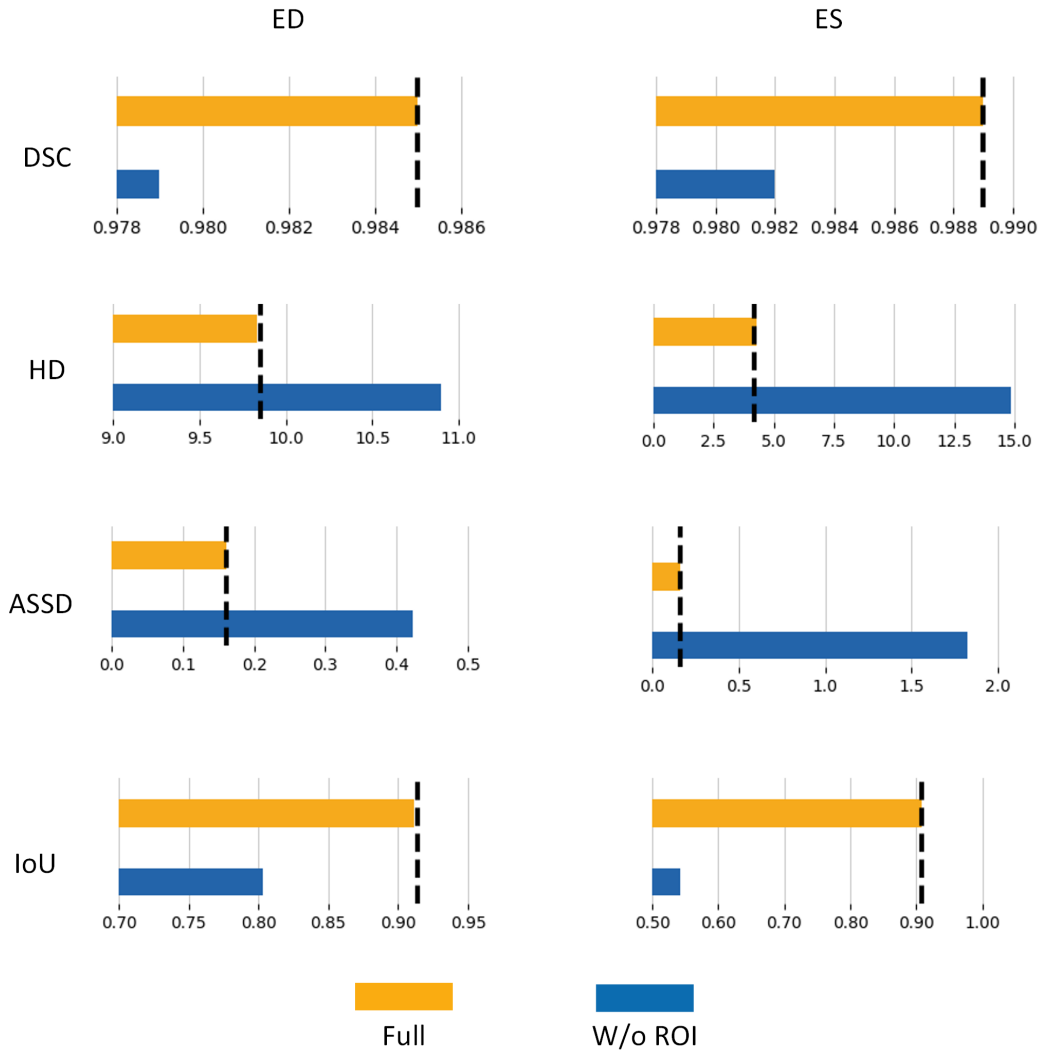

FIGURE 4.11: Effectiveness of the ROI extraction module (W/o ROI) in the proposed approach, as demonstrated by the ablation study. "Full" indicates the complete FC-TransNet framework, while "W/o" signifies the absence of this module.

Performance Comparison with the Base Models

Tables 4.7 and 4.8 present a quantitative comparison of FCTransNet and its transformer components during the ED and ES phases. The analysis includes DSC, HD, ASSD, and IoU metrics for three cardiac structures: LV, Myo, and RV. During the ED phase, FCTransNet consistently outperforms TransUNet, SwinUNet, and SegFormer across all evaluated metrics. Notably, it achieves superior DSC and IoU scores, along with lower HD and ASSD values, reflecting enhanced accuracy and precision in segmenting all cardiac structures. However, TransUNet surpasses FCTransNet in the ASSD for the LV class and the HD for the Myo class. In the ES phase, FCTransNet continues to surpass the base models, achieving better DSC and IoU scores, along with reduced HD and ASSD values.

Fusion method	Average				LV			Myo			RV		
	DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
TransUNet	0.983	15.120	0.163	0.911	0.963	3.000	**0.125**	**0.895**	16.309	0.160	0.958	16.095	0.145
SwinUNet	0.967	14.476	0.370	0.849	0.950	5.000	0.199	0.845	18.000	0.229	0.880	16.633	0.449
SegFormer	0.972	11.835	0.229	0.872	0.941	15.588	0.214	0.838	8.485	0.237	0.946	**5.000**	0.142
FCTransNet	**0.985**	**9.854**	**0.161**	**0.914**	**0.964**	**2.828**	0.129	0.892	**3.000**	**0.154**	**0.961**	15.095	**0.132**

TABLE 4.7: Comparison of FCTransNet's segmentation performance with the base transformer models during the ED phase.

Fusion method	Average				LV			Myo			RV		
	DSC	**HD**	**ASSD**	**IoU**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**
TransUNet	0.987	16.158	0.234	0.907	0.947	2.236	0.137	0.931	3.000	0.147	0.924	19.698	0.311
SwinUNet	0.979	12.389	0.525	0.861	0.934	4.123	0.166	0.888	3.606	0.243	0.880	10.915	0.798
SegFormer	0.974	7.779	0.315	0.852	0.931	3.606	0.180	0.892	4.472	0.261	0.855	13.038	0.464
FCTransNet	**0.989**	**4.172**	**0.165**	**0.908**	**0.950**	**2.236**	**0.122**	**0.933**	**3.000**	**0.147**	**0.930**	**7.280**	**0.226**

TABLE 4.8: Comparison of FCTransNet's segmentation performance with the base transformer models during the ES phase.

Figure 4.12 offers a detailed analysis of segmentation performance, comparing FCTransNet with the individual transformers (TransUNet, SwinUNet, and SegFormer) during both the ED and ES phases. FCTransNet achieves remarkable results, with average Dice coefficients reaching 0.985 for ED and 0.989 for ES. Additionally, it sustains low average HD values (9.854 for ED and 4.172 for ES) and ASSD values (0.161 for ED and 0.165 for ES). Moreover, FCTransNet demonstrates outstanding performance not only in overall metrics but also in precisely delineating individual cardiac structures, including the LV, Myo, and RV. The model achieves high Dice coefficients along with minimal HD and ASSD values, underscoring its accuracy in capturing these essential anatomical regions.

4.5 Discussion

In this chapter, we present FCTransNet, an advanced computer-aided diagnosis system designed for segmenting cardiac MRI scans. FCTransNet comprises three robust ViTs: TransUNet, SwinUNet, and SegFormer. The system is structured around two core components: the ROI extraction module and the fusion module. Initially, we deploy a specialized UNet architecture to precisely locate and extract ROI from cardiac cine MRI images. Following this, the fusion module combines the predictions from the transformers using our new image fusion technique called IWST. This technique incorporates weights that offer global and local perspectives, ensuring that the fusion process benefits from the overall context provided by the masks (global view) and the detailed class information of individual pixels and their neighbors (local view). This method enhances the adaptability and contextual accuracy of the image fusion, leading to improved precision in the final segmentation. FCTransNet demonstrates its effectiveness by consistently achieving high accuracy in segmenting cardiac cine MRI images, as demonstrated by our experimental results.

The results demonstrated that FCTransNet exhibited a notable superiority over existing methods for cardiac structure segmentation. Compared with transformer-based and other DL techniques (Figure 4.6), FCTransNet consistently demonstrated superior performance across critical metrics (Tables 4.1 and 4.2). The model achieved an impressive DSC of 0.985 and an exceptional mean IoU of 0.914 during the ED phase. This reflects a high degree of overlap between the predicted and actual segmentations, and it outperforms the competition. Furthermore, the high degree of precision demonstrated by FCTransNet is evidenced by its low HD95 value of 1 mm, which illustrates the accuracy with which it can capture cardiac structures. Collectively, these results establish FCTransNet as a leading solution for cardiac cine MRI segmentation, outperforming both transformer-based and other DL methods.

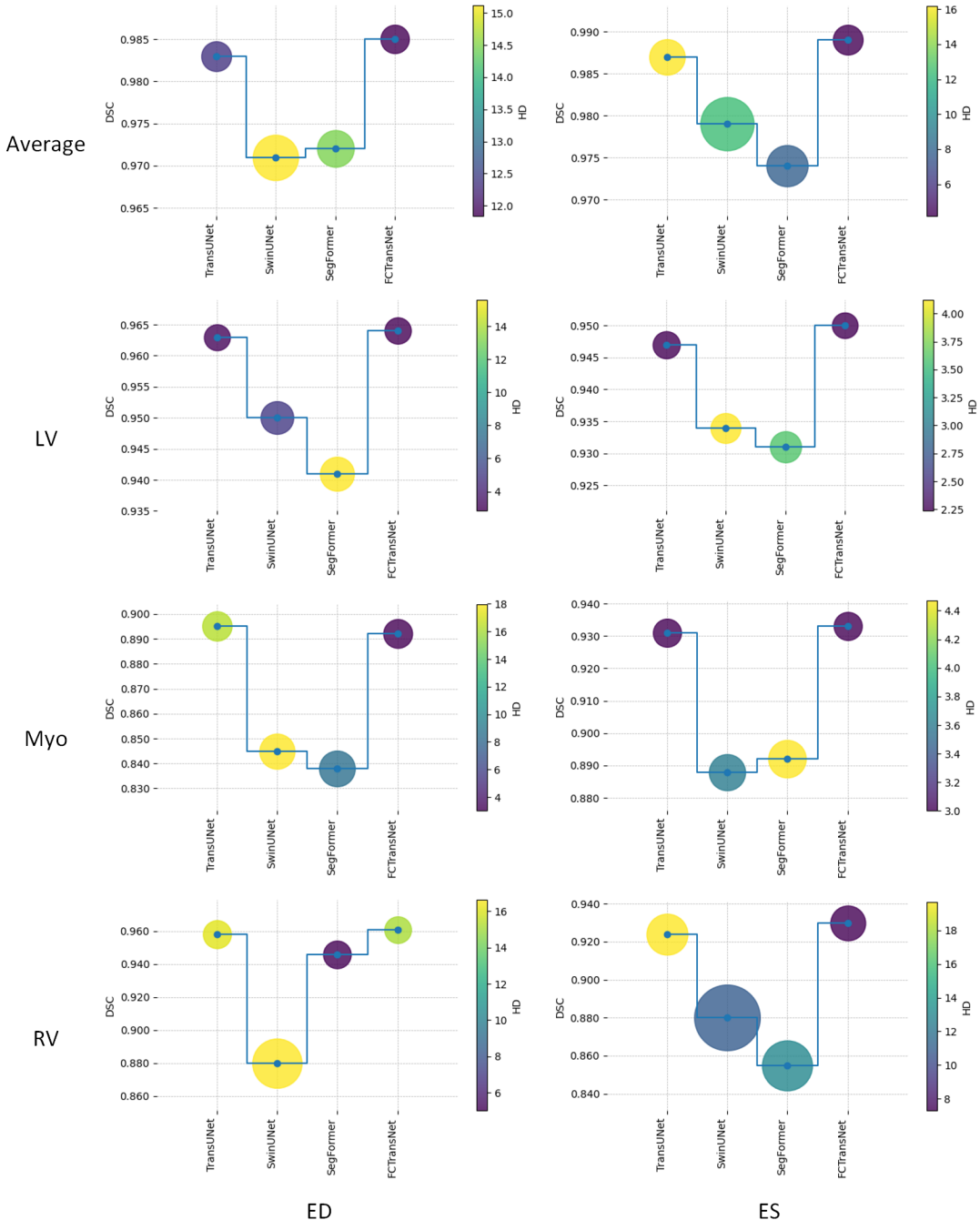

FIGURE 4.12: Segmentation performance of FCTransNet during the ED and ES phases. The method names are shown on the x-axis, and the DSC on the test dataset is plotted on the y-axis. The stair line represents the average DSC for each method, the circle size denotes the ASSD value, and the circle color indicates the HD value. FCTransNet demonstrates superior performance in DSC, ASSD, and HD compared to all individual transformers on the test dataset.

FCTransNet's superior segmentation accuracy, when compared to other transformer-based methods, is due to several critical factors. Many of these transformer-based models were trained directly on the ACDC dataset without incorporating essential preprocessing steps like ROI extraction or employing data augmentation techniques. Additionally, these models typically function as standalone systems, lacking the robustness achieved by ensemble methods like FCTransNet. In contrast, while some deep learning-based approaches have utilized preprocessing techniques, including ROI extraction, and have segmented the process into multiple stages [193, 195], FCTransNet still surpasses them. This success is due to the unique strengths of ViTs, specifically designed for medical image segmentation.

Tables 4.3 and 4.4 showcase the superiority of the IWST image fusion method compared to three advanced fusion techniques: averaging, weighted averaging, and voting. By incorporating global and local pixel details, IWST enhances segmentation accuracy and precision. The method consistently surpasses traditional techniques, as demonstrated by its higher precision scores. A visual illustration of these outcomes is presented in Figure 4.8, illustrating the qualitative segmentation results achieved by IWST.

The ROI location module's significance is demonstrated in Tables 4.5 and 4.6. This module notably improves the performance of both the base models and FCTransNet by decreasing computational load and enabling the model to concentrate on the most pertinent image regions, avoiding background areas that do not add valuable information. Focusing on the region of interest effectively mitigates the challenge of class imbalance in segmentation tasks, where the background often dominates as the largest class. This targeted approach not only enhances segmentation accuracy but also boosts computational efficiency. Figure 4.10 supports this by displaying the qualitative segmentation outcomes of the base models and FCTransNet on the original images from the ACDC dataset without ROI extraction. The critical role of the ROI extraction module is highlighted in Figure 4.11, which presents the improvements in DSC, HD, IoU, and ASSD metrics. Models utilizing the ROI extraction module achieve better results than those without it, a trend also observed in the FCTransNet approach.

The collaborative integration of the three ViTs leverages the unique strengths, complementarity, and diversity of each model. Tables 4.7 and 4.8 illustrate the comparative segmentation accuracy of each transformer and the combined FCTransNet approach. Combining these models leads to enhanced performance across all four evaluated metrics (Figure 4.12). Ultimately, the outstanding segmentation accuracy achieved by FCTransNet highlights the significance of a meticulously crafted architecture for cardiac image segmentation, affirming its superiority over transformer-based and other DL methods.

4.6 Conclusion

This chapter presented and examined FCTransNet, an ensemble computer-aided diagnosis approach specifically designed for segmenting cardiac anatomical structures, including the LV, Myo, and RV, in short-axis cine-MRI scans. The accurate segmentation of these structures is essential for the early detection and treatment of cardiovascular diseases.
Leveraging the advanced capabilities of state-of-the-art ViTs, our approach has demonstrated superior performance in cardiac structure segmentation, surpassing both ViT-based methods and other DL techniques. The study underscores the significance of a well-architected strategy for cardiac image segmentation and highlights the potential of ensemble techniques in enhancing medical imaging outcomes. The successful integration of the ROI extraction module, along with the novel image fusion technique, has led to notable improvements in segmentation accuracy. Specifically, the carefully crafted UNet architecture within the ROI module ensures precise localization and extraction of the region of interest, optimizing the delineation of cardiac structures.
By integrating the IWST, our approach effectively integrates the predictions from all three ViTs combining their collective strengths and addressing the challenges in cardiac cine MRI image segmentation. The fusion process benefits from the complementary and diverse capabilities of these models, resulting in enhanced segmentation accuracy and a higher level of precision in delineating cardiac structures. The success of our approach is evidenced by the performance evaluation on the ACDC test dataset, where FCTransNet achieves competitive results, particularly in the segmentation of the LV and Myo.
However, FCTransNet is not without limitations. Despite its promising results, the approach requires substantial computational resources, which may pose challenges in resource-constrained environments. Future research aims to address this limitation by exploring compression techniques that could reduce latency and computational costs, thereby enhancing the model's efficiency. This will form the basis of our next contribution, focusing on optimizing the performance of ViTs for broader application in clinical settings.

Chapter 5

Improved Two-stage Transfer Learning Approach for ViT-Based Myocardial Infarction Detection

5.1 Introduction

ViTs have emerged as a powerful alternative to CNNs due to their ability to capture images' long-range dependencies using attention mechanisms [197]. This capability allows ViTs to achieve remarkable performance in image classification [198] and segmentation tasks [9], particularly in computer vision. However, despite these advancements, DL techniques in medical imaging face several challenges. The need for large datasets to ensure model generalizability is a significant obstacle in the medical domain, where data collection is often constrained, and annotation requires specialized expertise from professionals like radiologists [199]. Additionally, training large DL models requires substantial computational resources, further complicating their deployment in medical applications. To address these limitations, TL has been introduced as a promising approach in medical image analysis [200]. TL enables the application of knowledge gained from a source domain to improve learning in a related target domain, thus avoiding the need for large labeled datasets in the target domain [201]. However, the success of TL largely depends on the alignment between the source and target domains. When the domains differ significantly, particularly in data distributions, the effectiveness of knowledge transfer diminishes, potentially impacting the model's performance [202]. Careful selection of the source domain is crucial to achieving successful outcomes, especially when working with limited labeled data in medical imaging.

This chapter introduces a two-stage transfer learning approach for detecting and diagnosing MI from CMR images, using ViT models as the backbone model. By leveraging a pretrained ViT model on a large, diverse cardiac dataset and fine-tuning it on a smaller, MI-specific dataset, the approach ensures strong domain alignment. This alignment enhances the model's ability to generalize well in diagnosing MI, highlighting the significance of shared features and patterns across domains to achieve optimal performance. The main contributions presented in this chapter are:

- An improved two-stage TL approach, combining network-based and instance-based techniques, is proposed to optimize the ViT model layers and leverage knowledge from a classification dataset, enhancing segmentation performance while reducing training time and data requirements.
- An effective framework for computer-aided diagnosis of MI disease is proposed and rigorously evaluated.
- An enhanced ViT model featuring a tailored weighted loss function, specifically designed to tackle class imbalance in segmentation, is explored as the core model to optimize the proposed two-stage TL approach.

The rest of the chapter is organized as follows: Section 5.2 presents an overview of recent related works on TL approaches for cardiac image segmentation. Section 5.3 details the proposed method. Section 5.4 describes the conducted experiments and the obtained results. Section 5.5 discusses the significance of these findings. Finally, Section 5.6 concludes with insights into potential future directions.

5.2 Related Works

Several recent studies in the literature have introduced TL methods for cardiac image segmentation tasks [203]. Chen et al. [204] proposed a CNN-based approach for segmenting the LV myocardium in porcine cardiac cine MRI. Their model was first trained on a public human cardiac MRI dataset with 9300 images and then fine-tuned by removing the softmax layer, followed by further training on 3600 porcine cardiac MRI scans. The final model achieved a dice score of 0.86. Another CNN model was employed for segmenting the left atrial LA cavity in 100 3D CMR volumes [205], where it was pretrained on the ImageNet dataset and fine-tuned using the MICCAI 2018 Atrial Segmentation challenge, resulting in a dice score of 0.92. Similarly, in [206], Serrano-Antón et al. applied a simple TL approach to their proposed UNet architecture for coronary artery segmentation from CT coronary angiography images, pretraining the model on

ImageNet and fine-tuning it on a private cardiac dataset.

Zhu et al. [207] explored the effectiveness of retraining a pretrained CNN model for myocardium segmentation in their study. The aim was to assess how well knowledge transferred from one MRI modality (T1-weighted) to two others (T2-weighted and extracellular volume quantification). The model was pretrained using 11,550 T1-weighted MRI images and tested on 1,525 T2-weighted and 1,525 EVQ images. Similarly, Ankenbrand et al. [208] investigated the impact of knowledge transfer across MRI modalities. Their CNN model was pretrained on the UK Biobank dataset for LV and Myo segmentation, then fine-tuned using the 2015 Data Science Bowl Challenge dataset. The approach resulted in dice scores of 0.90 for LV and 0.79 for myocardium segmentation.

In [209], an unsupervised domain adaptation-based TL technique was applied to a multi-scale CNN model for segmenting cardiac coronary angiography images. This process facilitated the knowledge transfer from a labeled to an unlabeled dataset. The application of this TL technique on the test set resulted in an increase in the model's average segmentation accuracy from 0.92 to 0.93. Similarly, Dong et al. [210] introduced a TL framework for whole heart segmentation, where knowledge was transferred from MRI to CT images using the MM-WHS dataset. Koehler et al. [211] proposed an unsupervised domain adaptation strategy to transfer knowledge between short-axis and axial cardiac MRI scans. They employed a UNet model, initially trained on short-axis images from sub-cohort 2 of the Tetralogy of Fallot dataset, which was subsequently fine-tuned on axial images from sub-cohort 1 of TOF and the ACDC dataset. The dice scores for the LV, RV, and Myo structures were 0.86, 0.77, and 0.65, respectively.

The use of TL to enhance segmentation performance is a common practice in current research on cardiac image segmentation. Still, the technique's contribution is rarely highlighted. A common approach involves using the ImageNet dataset as the source domain, despite its limited relevance to cardiac structures. Many of these studies overlook the importance of selecting a more appropriate source domain for cardiac-specific tasks. Furthermore, none of the current works in cardiac image segmentation utilize a classification cardiac dataset to pretrain the model's encoder [7]. This is critical, as pretraining with such a dataset could significantly improve the model's ability to capture the anatomical details of the heart, thereby enhancing segmentation performance.

While ViTs outperform CNNs in capturing long-range dependencies in computer vision [212], they have yet to be applied to TL in cardiac image segmentation. This chapter introduces a method that fills the gaps in the literature, showcasing exceptional performance on the test dataset. Unlike existing TL methods, the proposed approach capitalizes on training the ViT model's encoder with a cardiac dataset for classification tasks, followed by pretraining it with labeled images from the same dataset for segmentation tasks. Finally, the model is fine-tuned on a smaller, separate dataset.

This two-stage transfer learning strategy significantly improves the model's ability to generalize, regardless of the size of the target dataset.

5.3 Two-Stage Transfer Learning Framework

The proposed approach introduces a two-stage TL framework designed to enhance myocardial infarction segmentation in MRI images by leveraging the strengths of ViT models, specifically the TransUNet (Figure 5.1), which has shown promising results in our previous experiments. The method combines knowledge from both classification and segmentation tasks to improve overall performance.
The approach is structured in three main phases: preprocessing, pretraining, and finetuning. Initially, the preprocessing phase isolates the ROI from the cardiac MRI datasets. Following this, the pretraining phase, consisting of two stages, involves first pretraining the backbone of the TransUNet model on a cardiac classification dataset using instance-based and network-based transfer learning techniques. In the second stage, the TransUNet model, equipped with the pretrained backbone, is trained on a cardiac segmentation dataset through network-based transfer learning. Finally, in the finetuning phase, the pretrained TransUNet model undergoes fine-tuning on a dedicated MI diagnosis dataset, ensuring optimal adaptation to the specific task. This structured approach enables more accurate segmentation by building on the knowledge gained from multiple tasks and progressively refining the model.

5.3.1 Datasets and Preprocessing

5.3.1.1 ACDC Datasets

For the pretraining phase of the proposed framework, two datasets are utilized: the ACDC classification dataset and the ACDC segmentation dataset. The ACDC classification dataset categorizes images into five groups: healthy individuals, those with a history of myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and subjects with RV abnormalities. The ACDC segmentation dataset provides manual annotation of those images, where a single expert performed segmentation on 2D cine slices during the ES and ED phases, outlining the RV, Myo, and LV. The ACDC classification and segmentation datasets comprise the same number of MRI images.

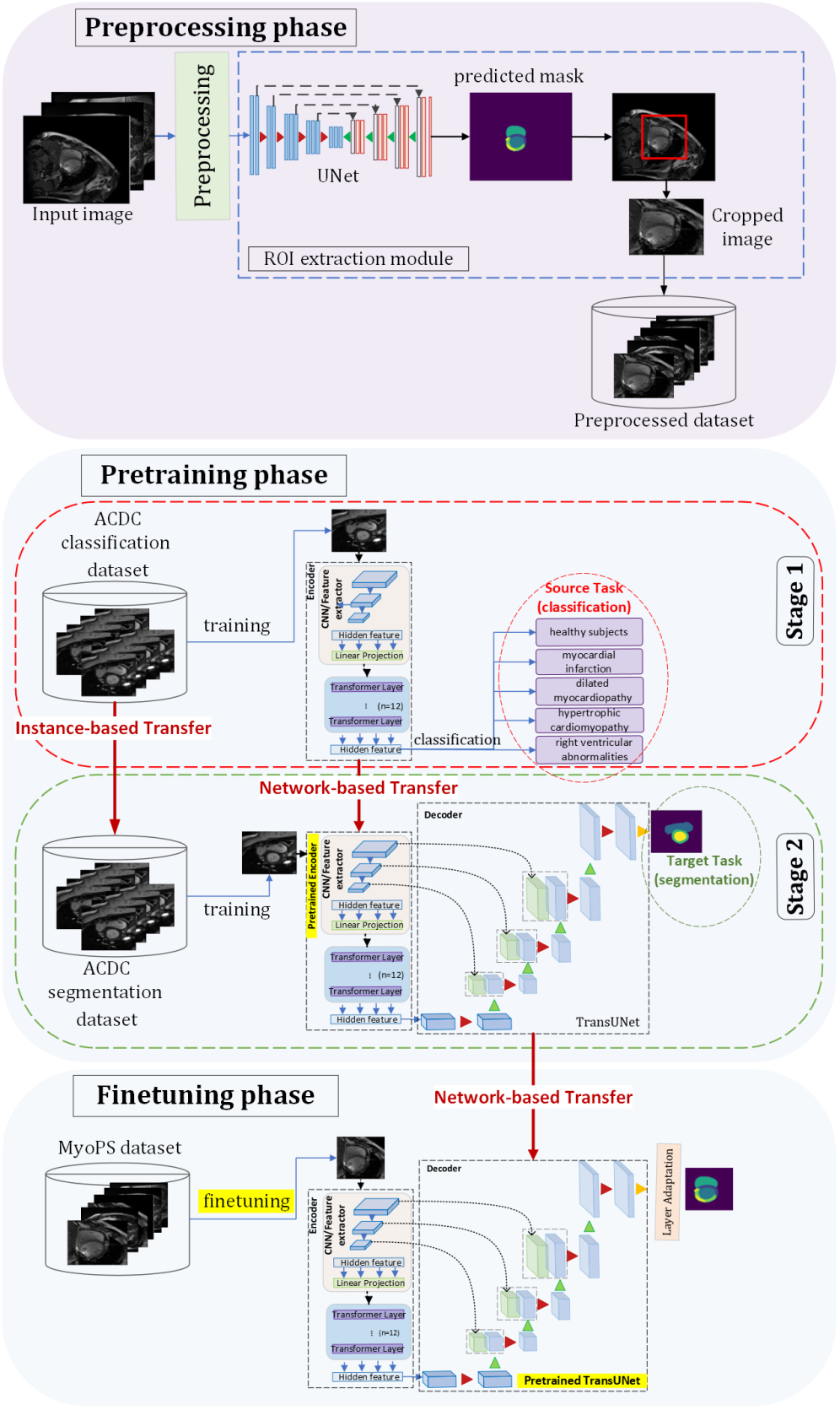

FIGURE 5.1: Overview of the proposed two-stage TL framework for MRI MI segmentation.

5.3.1.2 MyoPS Dataset

The MyoPS dataset consists of 45 multi-sequence CMR volumes collected from the Myocardial Pathology Segmentation Challenge 2020 (MyoPS 2020) [213]. These volumes have an average resolution of $482 \times 479 \times 4$ pixels and include three CMR sequences: bSSFP, LGE, and T2. All cases were aligned and resampled to a unified spatial resolution using the MvMM technique [63] to ensure consistency in the analysis. The dataset is divided into 25 training cases and 20 test cases, with the training subset containing 306 annotated slices and the test set containing 216 slices. Expert annotations mark various regions of interest, including LV blood pool, RV blood pool, LV normal myocardium, LV myocardial edema, and LV myocardial scars. The 306 slices from the training set are split into training and validation subsets at an 80:20 ratio.

All preprocessing steps were applied uniformly to the segmentation datasets, following the same procedure as in our earlier work. These steps involved resizing the images to a resolution of 224×224 pixels, normalizing pixel intensities, utilizing a U-Net model for ROI extraction, and applying data augmentation techniques to increase the variability and resilience of the training data. Figure 5.2 shows an example of 2D slices taken from the same CMR volume, presented both before and after ROI extraction, where the top row illustrates the original image and the bottom row displays the corresponding extracted ROI. The ROI extraction process was mainly dedicated to the segmentation datasets, as it leverages the U-Net architecture, which takes an input image and produces a corresponding segmentation mask. In contrast, the ACDC classification dataset was created from the preprocessed ACDC segmentation dataset by isolating only the input images.

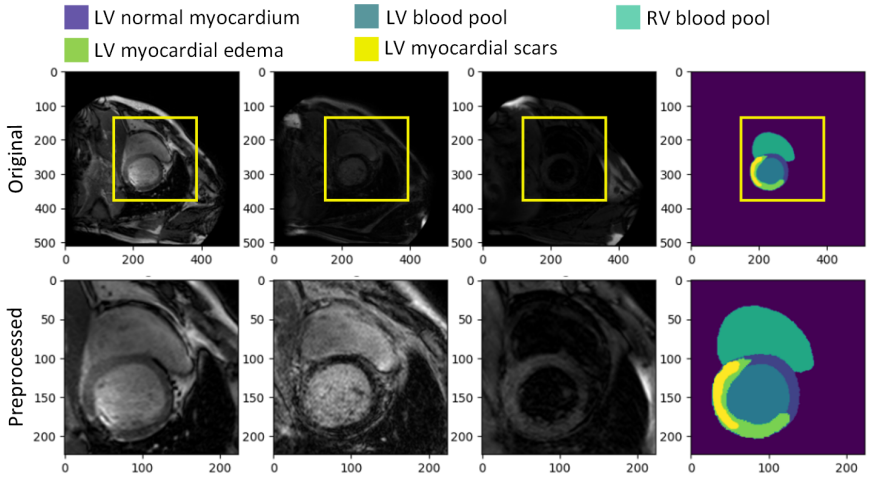

FIGURE 5.2: Example of 2D slices taken from the same CMR volume, shown before and after ROI extraction. The top row illustrates the original image, while the bottom row presents the corresponding extracted ROI.

5.3.2 Pretraining Phase

TL refers to utilizing knowledge acquired from a source domain D_s in a specific task T_s to address a target domain D_t and solve a related target task T_t. In a TL setting:

- The source domain $D_s = \{X_s, P_s(X_s)\}$ consists of a feature space X_s and its marginal probability distribution $P_s(X_s)$.
- The source task $T_s = \{Y_s, P_s(Y_s|X_s)\}$ includes output labels Y_s and the conditional probability distribution $P_s(Y_s|X_s)$.
- The target domain $D_t = \{X_t, P_t(X_t)\}$ contains the feature space X_t and its associated marginal probability distribution $P_t(X_t)$.
- The target task $T_t = \{Y_t, P_t(Y_t|X_t)\}$ includes the output labels Y_t and the conditional probability distribution $P_t(Y_t|X_t)$.
- The predictive function $f_T(\cdot)$ represents the model.

In this context, the predictive function $f_T(\cdot)$, which represents the TransUNet model for myocardial infarction detection, is enhanced by identifying and transferring important hidden knowledge from the source domain D_s to the target domain D_t to boost the model's performance.
The proposed TL approach in this work integrates two types of TL techniques, namely, *instance-based transfer* and *network-based transfer*. The pretraining phase is composed of two main stages; the first stage involves training the backbone of the TransUNet model on the ACDC classification dataset, exposing the model to various cardiac conditions. This initial training imparts essential cardiac feature knowledge crucial for classification. In the second stage, the TransUNet model, equipped with the pretrained backbone, is further trained on the ACDC segmentation dataset. This phase refines the model's understanding of cardiac structures and their spatial relationships, critical for precise segmentation. The approach is designed to enhance myocardial infarction detection by integrating both classification and segmentation capabilities.

5.3.2.1 Stage 1: Pretraining on the Classification Task

In this phase, the initial pretraining involves training the backbone of the TransUNet model using the ACDC classification dataset. The goal is to capture general features pertinent to cardiac conditions through the classification process. By considering the second stage of the approach as a target domain, this stage can be formulated as follows:

- **Source domain (D_s)**: ACDC classification dataset
- **Source task (T_s)**: Classification task
- **Target domain (D'_s)**: ACDC segmentation dataset
- **Target task (T'_s)**: Segmentation task

The primary objective of this phase is to apply the features acquired from the classification task T_s using the ACDC classification dataset D_s to the segmentation task T'_s using the ACDC segmentation dataset D'_s, as illustrated in Figure 5.3. This adaptation is facilitated through a combined TL approach that merges instance-based TL with network-based TL. Instance-based TL employs the knowledge derived from the samples and their distribution in D_s, while network-based TL entails the utilization of model weights and parameters from T_s, integrating the pretrained backbone within the TransUNet model to accommodate T'_s. The objective of synthesizing these TL methods is to enhance the model's accuracy in detecting cardiac structures during the segmentation task by leveraging the features learned from the classification task, as follows:

$$f\left((D_s \to T_s) \times (D'_s \to T'_s)\right) \to f\left(((D_s \to D'_s) \to T'_s) \to (T_s \to T'_s)\right) \tag{5.1}$$

5.3.2.2 Stage 2: Pretraining on the Segmentation Task

After pretraining the backbone of the TransUNet model on the ACDC classification dataset, the whole TransUNet model is then trained on the ACDC segmentation dataset. This process adapts the model to the segmentation task by utilizing both the input images and their associated masks. The procedure can be expressed as follows:

- **Source domain (D'_s)**: ACDC segmentation dataset (Refined source domain)
- **Source task (T'_s)**: Segmentation task on ACDC dataset
- **Target domain (D_t)**: MyoPS dataset
- **Target task (T_t)**: Segmentation task on MyoPS dataset

In this stage, network-based TL is utilized to fine-tune the TransUNet model using the ACDC segmentation dataset D'_s. This approach is mainly proposed to transfer knowledge gained during pretraining on the source classification task T'_s and adjust it for the specific needs of the target segmentation task T_t. By applying network-based TL, the model can adapt more efficiently to the segmentation task on D_t, leveraging the features

it learned during pretraining on D'_s. This reflects the transition of the model from performing the classification task T'_s on D'_s to accurately handling the segmentation task T_t within the target domain D_t. The objective is to improve the model's precision in segmenting cardiac structures within the MyoPS dataset. The second stage is represented as follows:

$$f\left((D'_s \to T'_s) \times (D_t \to T_t)\right) \to f\left(D_t \to \left(T'_s \to T_t\right)\right) \tag{5.2}$$

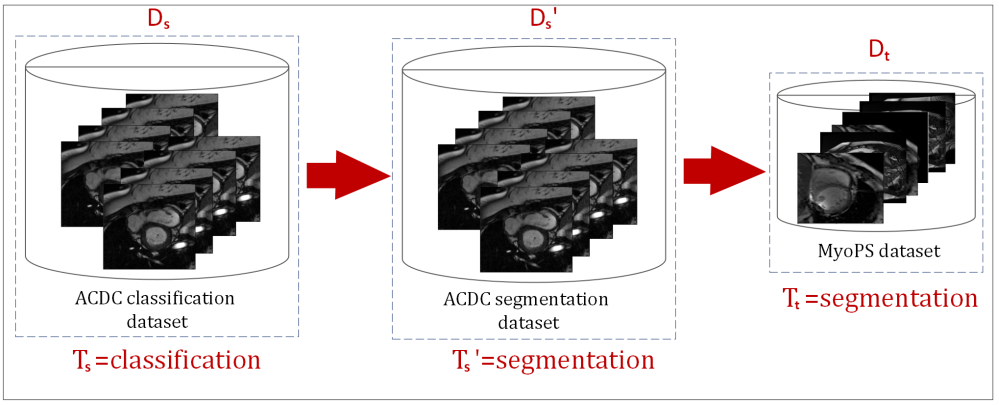

FIGURE 5.3: An overview of the two-stage TL approach.

5.3.3 Finetuning Phase

After pretraining the TransUNet model and its backbone on two extensive and heterogeneous datasets (ACDC classification and segmentation datasets), the model is subsequently finetuned on the smaller cardiac MyoPS dataset. This finetuning stage is a crucial aspect of machine learning, as it enables the transfer of the general knowledge acquired during the pretraining phase to align with the specific requirements of the target task. This involves refining the model by adapting its features and parameters to suit the complexities of myocardial pathology segmentation. Finetuning is often seen as a subtle but effective approach, ranging from minor parameter adjustments to substantial modifications in the model's architecture. The goal is to improve the model's accuracy and performance for the MyoPS dataset by leveraging TL, ensuring its predictions are well-suited to the intricate nature of myocardial pathology segmentation.

During the finetuning phase of the TransUNet model, all the parameters are adjusted to align with the specific segmentation task. This process involves updating the entire model, from the backbone to the output layer. Before parameter adjustment, the final layer is modified to meet the task's requirements. This step includes reorganizing the encoded feature map to restore spatial structure and applying a 1×1 convolution to match the number of output channels with the target classes. By finetuning the entire model and customizing the final layer, the TransUNet model effectively adapts its

features and parameters to the complexities of the segmentation task, improving both precision and performance.

5.3.4 Weighted Loss Function

The training of segmentation models is susceptible to the effects of imbalanced medical datasets, which can lead to suboptimal performance on both underrepresented classes and the background. To address this challenge, several loss functions have been introduced, including dice loss and cross-entropy loss. For all experiments, we employed a hybrid loss function that combines both cross-entropy and Dice loss. However, in the case of the MyoPS dataset, we encountered a significant imbalance not only between the target region and the background but also among different classes within the target region itself. This increased the complexity of the problem, emphasizing the necessity of a weighted loss function capable of effectively addressing these imbalances.
Class weights are introduced to modify the loss function, ensuring that the model gives appropriate emphasis to each class. These weights are computed based on the inverse frequency of each class. In particular, the weight for a given class i is calculated as:

$$W_i = \frac{N}{C \times S_i} \tag{5.3}$$

Where N represents the total sample count for all classes, C is the total number of classes, and S_i denotes the number of samples for class i.
Weighted Dice Loss is an adaptation of the standard Dice Loss that integrates class weights to address issues of class imbalance. The weighted Dice Loss function is formulated as:

$$\mathcal{L}_{\text{WDice}}(G, P) = \sum_{i=0}^{K-1} W_i \times (1 - \mathcal{L}_{\text{Dice}}(G, P)) \tag{5.4}$$

Where G and P represent the ground truth and the predicted segmentation mask, respectively. $\mathcal{L}_{\text{Dice}}$ is the standard Dice loss function.

The hybrid loss function is defined as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{WDice}} \tag{5.5}$$

This hybrid loss function has been designed to enhance the model's performance on imbalanced datasets. The incorporation of cross-entropy loss facilitates precise pixel-level classification, whereas the weighted Dice loss guarantees that the segmentation accurately captures the structural characteristics of target regions, even when those regions are under-represented.

5.4 Experimental Results

This section presents and analyzes the experimental results obtained from the proposed approach. First, the segmentation outcomes are thoroughly detailed, both quantitatively and qualitatively. Then, ablation studies are conducted to assess the significance and impact of each phase within the approach. Additionally, an in-depth analysis is performed to evaluate how different factors, such as the choice of the loss function, freezing weights during the finetuning phase, and the generalizability of the model compared to pretraining on the ImageNet dataset, influence the overall performance. Finally, the results are compared with previous work on the MyoPS dataset. The models' segmentation performance is evaluated using four metrics: DSC, IoU, HD, and ASSD. The experiments follow the same setup as described in the earlier contributions.
To validate the effectiveness of our approach, we conducted a detailed comparison of our model's segmentation results with those produced by leading state-of-the-art models. Leveraging the TransUNet architecture, which combines CNN and Transformer blocks, we selected well-established models from both categories. Specifically, we included CNN-based models such as UNet and Attention-Unet, alongside Transformer-based models like SwinUNet and Segformer, which are mainly dedicated to medical image segmentation.

5.4.1 Segmentation Results

In this section, we conduct a comprehensive evaluation of the proposed method by comparing its performance to the previously mentioned DL models that are widely employed in the medical image segmentation field.

5.4.1.1 Quantitative Evaluation

Table 5.1 presents the segmentation results of the proposed approach against the different DL models used for medical image segmentation, focusing on Myo, LV, RV, edema, scar, and combined edema and scar segmentation. Our proposed method demonstrates superior performance across multiple metrics, including the DSC, HD, ASSD, and IoU. In terms of overall performance, our approach achieves the highest Dsc for Myo segmentation at 0.771 and 0.912 for the LV, significantly surpassing other models such as UNet, AttUNet, SwinUNet, and SegFormer. This improvement indicates that our method effectively captures the intricacies of myocardial structures, which is crucial for accurate clinical assessments. For LV segmentation, our model again outperforms with a Dsc of 0.912, surpassing AttUNet (0.891) and SwinUNet (0.859), reinforcing its capability to

delineate the left ventricle's boundaries.

The RV segmentation also shows high results, with our method achieving a Dsc of 0.899, higher than SwinUNet (0.782) and SegFormer (0.688). This consistent performance across multiple cardiac structures highlights the robustness of our approach. Furthermore, in edema segmentation, our model achieves a Dsc of 0.758, substantially higher than all other models, demonstrating its robustness in accurately detecting edema. Additionally, for scar segmentation, our approach maintains a high Dsc of 0.729, compared to AttUNet (0.446) and SegFormer (0.344).

The performance of our model on combined edema and scar segmentation is also noteworthy, achieving a DSC of 0.744. This result suggests that our method effectively integrates information from both conditions, improving overall accuracy. When examining the distance metrics, the HD and ASSD values reflect our method's efficiency in minimizing segmentation errors. For instance, the HD for the myocardium is notably low at 14.353, compared to AttUNet (35.398) and SwinUNet (34.029). Two comparison charts for Dsc values related to edema, scar, and edema+scar can be seen in Figure 5.4.

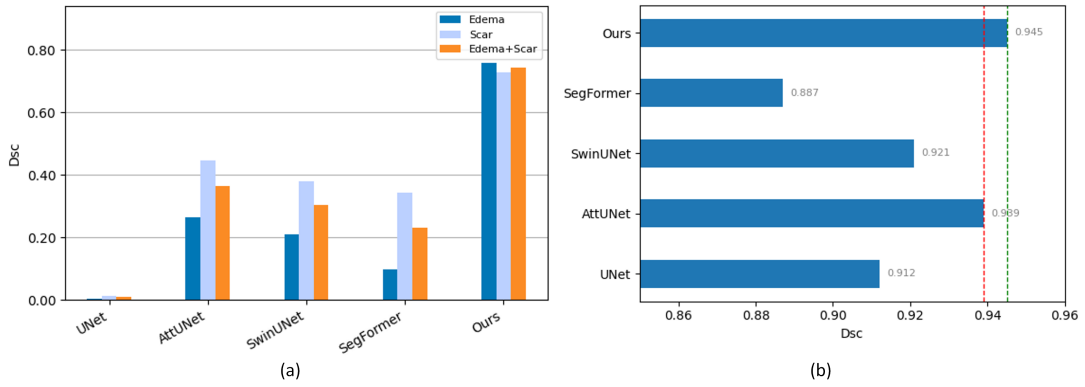

FIGURE 5.4: Bar chart comparison of different models. (a) DSC for edema, scar, and edema+scar, (b) Average DSC. The green and red lines in (b) show the average DSC score of our approach and the second best average DSC scor.

5.4.1.2 Qualitative Evaluation

Figure 5.7 presents a qualitative comparison of segmentation outcomes from various CMR images, illustrating results from the input image, ground truth, and segmentation models, including UNet, AttUNet, SwinUNet, SegFormer, and 2-TLViT. The regions of interest are colored with blue, green, red, yellow, and azure, indicating the LV normal myocardium, LV blood pool, RV blood pool, LV myocardial edema, and LV myocardial scars, respectively. Notably, 2-TLViT achieves consistent segmentation across all five

slices, accurately identifying the target regions. In contrast, the other models demonstrate inconsistencies in their segmentation, frequently failing to capture the target regions accurately. These errors are reflected in blank spaces within the segmented areas or the misclassification of background regions as part of the cardiac structures, highlighted by yellow rectangles. Such inconsistencies demonstrate the limitations of these models in handling complex cardiac features. Notably, these visual discrepancies are consistent with the quantitative results obtained, further validating the superior performance and robustness of 2-TLViT in delivering precise segmentation.

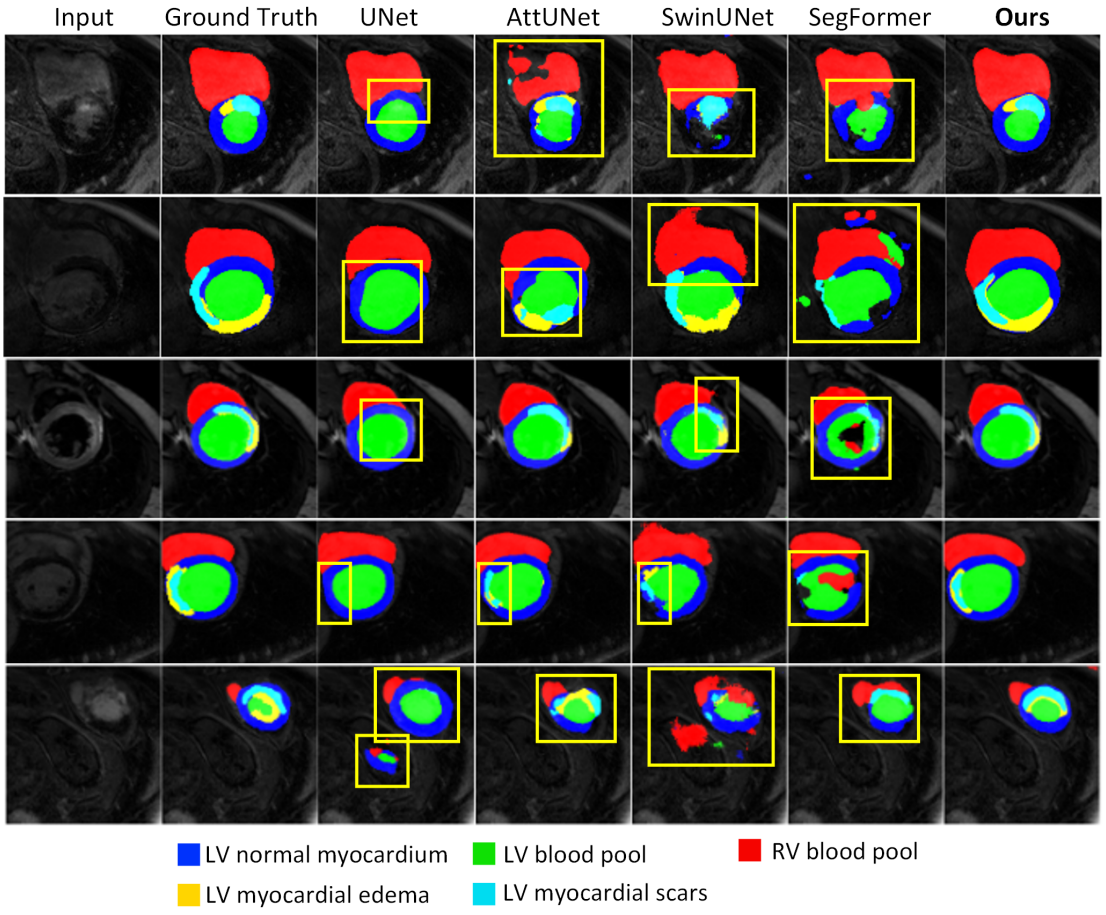

FIGURE 5.5: Segmentation results visualization by different state-of-the-art segmentation models compared to 2-TLViT.

5.4.1.3 Ablation Study

To assess the effectiveness of the two-stage transfer learning approach, we conducted a series of ablation experiments using the TransUNet architecture as the baseline. The evaluation considered various configurations: the baseline TransUNet without pretraining, pretraining the model's backbone on the ACDC classification dataset (`ACDC_class`),

pretraining the full model on the ACDC segmentation dataset (`ACDC_seg`), and pretraining both the backbone and the full model on the `ACDC_class` and `ACDC_seg` datasets, respectively. By varying the pretraining conditions, we aimed to isolate the individual contributions of each transfer learning stage and their combined effect on model performance. This analysis provided insights into the advantages of classification-based versus segmentation-based pretraining, highlighting the improved performance and robustness of the proposed method, as shown in Table 5.2.

In addition, to further evaluate the impact of the second pretraining phase, we pretrained only the backbone of the TransUNet model on the `ACDC_class` dataset and then fine-tuned it on the MyoPS dataset. We also tested the effectiveness of the first pretraining phase by pretraining the full TransUNet model on the `ACDC_seg` dataset without training the backbone. The results, presented in Table 5.2, demonstrate the performance gains over the baseline TransUNet model. Specifically, the TransUNet model pretrained on `ACDC_class` showed notable improvements, with a DSC of 0.882 for the LV and 0.837 for the RV, as well as significant reductions in HD and ASSD across most structures, indicating better boundary accuracy.

The model pretrained on `ACDC_seg` further enhanced performance, achieving a DSC of 0.893 for the LV and 0.853 for the RV. Additionally, it showed improved segmentation of scar tissue (DSC = 0.469) and combined edema and scar regions (DSC = 0.361).

Finally, the approach combining both pretraining phases resulted in the highest DSC scores across all regions. It significantly outperformed all other configurations, achieving DSC values of 0.771 for Myo, 0.912 for LV, 0.899 for RV, 0.758 for edema, 0.729 for scar, and 0.744 for the combined edema and scar regions. Moreover, it achieved substantial reductions in HD and ASSD, demonstrating superior segmentation accuracy and robustness.

The findings are presented in Figure 5.6, which displays the qualitative outcomes of the ablation experiments. This figure demonstrates noticeable enhancements in segmentation accuracy as each pretraining phase is applied. Additionally, Figure 5.7 features a line graph showing the average DSC scores, along with those for scar, edema, and their combined regions, highlighting the substantial improvements achieved by our proposed method.

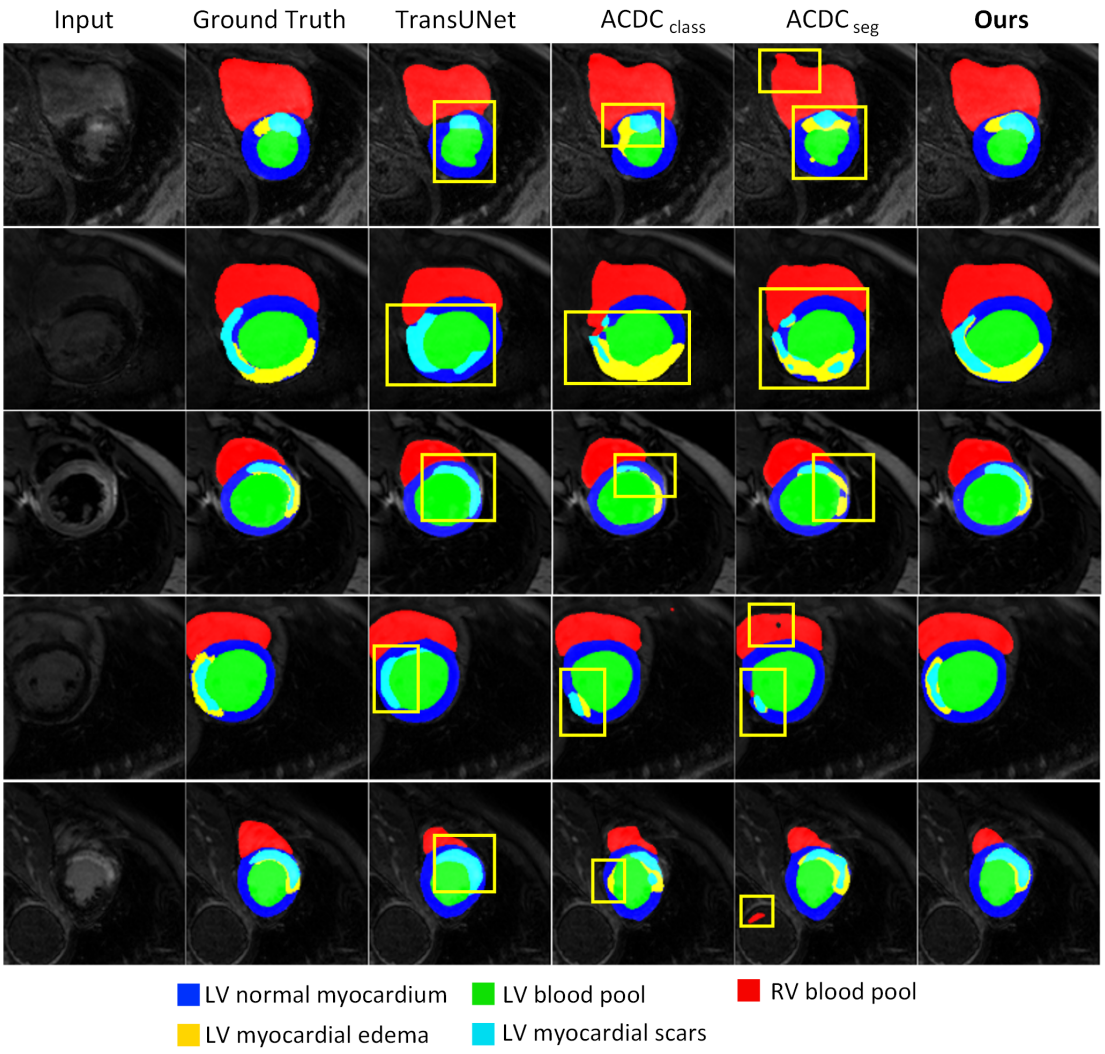

FIGURE 5.6: Segmentation results visualization of the ablation study.

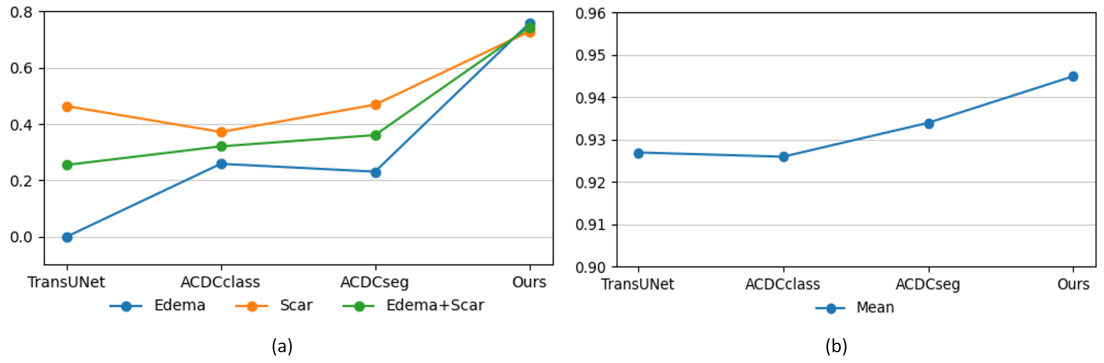

FIGURE 5.7: Line chart illustrating the DSC scores from the ablation experiments. (a) DSC for edema, scar, and edema+scar; (b) Average DSC.

Model	**Myo**			**LV**			**RV**			**Edema**			**Scar**			**E+S**	**Average**	
	DSC	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**IoU**	**DSC**
UNet	0.680	28.792	0.797	0.835	25.942	0.647	0.805	72.505	1.070	0.004	51.088	16.736	0.015	41.629	0.479	0.912	0.479	0.912
AttUNet	0.728	35.398	0.626	0.891	32.000	0.352	0.823	76.085	0.897	0.266	27.386	2.995	0.446	32.031	0.581	0.939	0.581	0.939
SwinUNet	0.690	34.029	0.697	0.859	111.400	0.386	0.782	56.868	0.823	0.209	29.017	4.062	0.380	24.228	0.539	0.921	0.539	0.921
SegFormer	0.668	65.085	1.234	0.758	79.612	2.227	0.688	106.836	2.259	0.098	77.208	9.908	0.344	81.062	0.473	0.887	0.473	0.887
Ours	0.771	14.353	0.439	0.912	12.207	0.245	0.899	50.843	0.237	0.758	20.712	2.279	0.729	29.223	0.643	0.945	0.643	0.945

TABLE 5.1: Quantitative results comparison with state-of-the-art segmentation models.

Model	**Myo**			**LV**			**RV**			**Edema**			**Scar**			**E+S**	**Average**	
	DSC	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**IoU**	**DSC**
UNet	0.706	39.235	0.634	0.800	31.613	0.398	0.801	57.911	0.551	0.000	1845.0	1845.0	0.364	30.712	0.507	0.917	0.507	0.917
AttUNet	0.707	34.176	0.634	0.882	19.000	0.345	0.837	74.229	0.694	0.259	38.013	3.553	0.372	44.283	0.567	0.926	0.567	0.926
SwinUNet	0.734	34.670	0.586	0.893	33.645	0.373	0.853	69.007	0.437	0.231	31.081	3.597	0.469	32.573	0.590	0.934	0.590	0.934
Ours	0.771	14.353	0.439	0.912	12.207	0.245	0.899	50.843	0.237	0.758	20.712	2.279	0.729	29.223	0.643	0.945	0.643	0.945

TABLE 5.2: Quantitative results comparison with state-of-the-art segmentation models.

5.4.1.4 Analysis Experiments

In this study, we examine the impact of various factors on the performance of 2-TLViT. We focus on the role of the loss function and the difference between freezing and unfreezing weights during the finetuning process. Furthermore, we explore the generalization capability of our approach by comparing its performance after pretraining on the ACDC datasets versus pretraining on the ImageNet dataset. This analysis aims to offer a detailed understanding of how these factors influence the efficiency and reliability of 2- TLViT.

Loss Function

To evaluate the effect of different loss functions, we performed experiments using both weighted and non-weighted loss functions. The class weights were calculated based on the class distribution of the dataset, as shown in Table 5.3. This weighting strategy aimed to address the class imbalance by giving more weight to the underrepresented classes, such as edema and scar tissue. Table 5.3 shows the class distribution along with the class weights applied in the weighted loss function. The Myo, LV, and RV classes are balanced, while the edema and scar classes are significantly underrepresented. To correct for this imbalance, we assigned higher weights to the edema and scar classes, with weights of 13.6 and 11.05, respectively, compared to 3.4 for Myo and 0.21 for the background.
Table 5.4 shows the performance of our method using non-weighted and weighted loss functions. The results indicate that the weighted loss function (2-TLViT) leads to a significant improvement in the DSC values for all structures. Specifically, the DSC for edema increased from 0.316 to 0.758 and for scar from 0.517 to 0.729. In addition, improvements in the HD and ASSD metrics indicate more accurate boundary delineation. The average DSC score increased from 0.937 to 0.945, demonstrating that the weighted loss function effectively addresses class imbalance.

Finetuning Phase

During the finetuning process, we tested both frozen and unfrozen weight configurations to assess the impact of TL on model performance. As illustrated in Table 5.5, the approach with unfrozen weights consistently achieved better results compared to the configuration of the frozen weight. For instance, the DSC for Myo improved from 0.765 to 0.771, and for LV from 0.906 to 0.912. These improvements suggest that allowing

the model to further adjust during the finetuning phase enhances its performance. Additionally, both HD and ASSD metrics showed favourable changes, reinforcing the idea that unfreezing the weights during finetuning leads to more accurate segmentation.

Generalizability

To evaluate the generalizability of our method, we compared the performance of our model when pretrained on the ACDC datasets to when pretrained on the ImageNet dataset. This comparison allows us to understand the benefits of domain-specific pretraining versus more general, broad-domain pretraining. As shown in Table 5.6, the model pretrained on the ACDC datasets consistently outperforms the model pretrained on ImageNet in all evaluation metrics. Specifically, the DSC for Myo improved from 0.698 to 0.771, and for Edema from 0.132 to 0.758. These significant performances highlight the advantage of using domain-specific data for pretraining, which allows the model to capture the unique features of medical images more effectively.

Class	**Myo**	**LV**	**RV**	**Edema**	**Scar**	**Background**
Class distribution	540480	562989	675803	135302	166545	8957601
Class weight	3.4	3.27	2.72	13.6	11.05	0.21

TABLE 5.3: Class distribution and corresponding class weights for the weighted loss function.

Model	Myo			LV			RV			Edema			Scar			E+S	Average	
	DSC	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**IoU**	**DSC**
Non-weighted	0.731	28.071	0.542	0.896	25.923	0.303	0.857	49.588	0.383	0.316	29.172	2.932	0.517	26.420	1.784	0.426	0.608	0.937
Weighted	0.771	14.353	0.439	0.912	12.207	0.245	0.899	50.843	0.237	0.758	20.712	2.279	0.729	29.223	2.055	0.744	0.643	0.945

Table 5.4: Performance comparison between weighted and non-weighted loss functions.

Model	Myo			LV			RV			Edema			Scar			E+S	Average	
	DSC	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**IoU**	**DSC**
Freeze	0.765	23.173	0.479	0.906	15.232	0.268	0.874	26.981	0.327	0.285	25.495	2.867	0.524	29.223	2.001	0.417	0.620	0.944
Unfreeze	0.771	14.353	0.439	0.912	12.207	0.245	0.899	50.843	0.237	0.758	20.712	2.279	0.729	29.223	2.055	0.744	0.643	0.945

Table 5.5: Performance comparison between freezing and unfreezing weights during the finetuning phase.

Model	Myo			LV			RV			Edema			Scar			E+S	Average	
	DSC	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**HD**	**ASSD**	**DSC**	**IoU**	**DSC**
ImageNet	0.698	30.232	0.733	0.867	22.023	0.449	0.792	59.008	0.948	0.132	29.257	4.560	0.354	29.069	3.299	0.254	0.534	0.923
ACDC	0.771	14.353	0.439	0.912	12.207	0.245	0.899	50.843	0.237	0.758	20.712	2.279	0.729	29.223	2.055	0.744	0.643	0.945

Table 5.6: Performance comparison between pretraining on ImageNet dataset and ACDC datasets.

5.4.1.5 Comparative Results

To assess the effectiveness of the proposed method, we conducted a comparative evaluation of the DSC scores for edema, scar, and the combined edema+scar, utilizing various techniques previously applied to the MyoPS dataset. Table 5.7 provides a comparison of the DSC scores for edema, scar, and the combined edema+scar across different studies on the MyoPS dataset, showcasing the performance of our approach relative to previous research.

For edema segmentation, 2-TLViT achieved a DSC score of 0.758, which is notably higher than those reported in prior studies. Although this score is slightly lower than the highest value of 0.767 reported by [214], it still demonstrates significant improvement compared to earlier methods.

Regarding scar segmentation, our approach delivered a substantial improvement, reaching a DSC score of 0.729. This score exceeds those from previous studies, including a slight improvement over the 0.719 reported by [214], which further highlights the robustness and efficacy of our method in segmenting scar tissue.

For the combined edema+scar segmentation, our method achieved a DSC score of 0.744, the highest among all the methods compared. The results demonstrate that our approach performs competitively and, in most cases, surpasses existing methods in segmenting both edema and scar. This highlights the effectiveness of our technique in accurately identifying these important regions within the MyoPS dataset, suggesting its potential for enhancing clinical applications in myocardial pathology segmentation.

Method	**year**	**Edema**	**Scar**	**E+S**
[215]	2020	0.731	0.672	0.702
[216]	2023	0.742	0.661	0.702
[217]	2023	0.735	0.678	0.707
[214]	2024	**0.767**	0.719	0.743
Ours	2024	0.758	**0.729**	**0.744**

TABLE 5.7: Comparison of Edema and Scar DSC scores with previous studies on the MyoPS dataset.

5.5 Discussion

In this study, we present a two-stage TL approach for the segmentation of MI. The fundamental concept of this method is to integrate the advantages of both classification-based and segmentation-based pretraining in order to optimize the performance of the segmentation model. The initial stage entails finetuning a pretrained model using the ACDC classification dataset, thereby enabling it to extract general features of relevance

to cardiac anatomy. This stage is crucial for enabling the model to acquire a diverse set of discriminative features that are broadly applicable to various segmentation tasks. In the second stage, the model is retrained on the ACDC segmentation dataset to adapt the general features for the specific task of segmenting cardiac tissues by learning to define the boundaries of these structures accurately. Finally, the model undergoes finetuning on the MyoPS dataset, which focuses on myocardial pathology segmentation. This task-specific adjustment further refines the model, enabling it to capture the detailed characteristics of myocardial infarctions and significantly improving segmentation performance.

The results from the ablation experiments (Table 5.2) indicate that combining both pretraining phases leads to a greater improvement in model performance compared to using each phase individually. The classification-based pretraining phase equips the model with generalized features that are useful for distinguishing various cardiac structures. These features provide a solid foundation for the subsequent segmentation pretraining phase. During this second phase, the model refines the previously learned features, tailoring them to the specific task of myocardial infarction segmentation, which ultimately enhances boundary precision and segmentation accuracy (Figure 5.7).

The results of our analysis, presented in Tables 5.3 to 5.6, further confirm the effectiveness of our proposed method. For example, Table 6 shows that the configuration using a weighted loss function (our approach) consistently outperforms the non-weighted counterpart across all metrics, including DSC, HD, and ASSD. This demonstrates the crucial role of incorporating class weights to address the class imbalance commonly found in medical imaging datasets. The enhanced performance in segmenting myocardial edema and scars is particularly significant, as these regions are critical for diagnosing and monitoring MI.

Additionally, Table 5.5 reveals that unfreezing the weights during the finetuning phase leads to better outcomes compared to keeping the weights frozen. This highlights the value of allowing the model to adjust and fine-tune its parameters during the finetuning phase, improving its ability to learn the specific features of the target dataset. Our approach, which includes unfreezing the weights, achieves higher DSC scores and lower HD and ASSD values, indicating more precise and dependable segmentation. Table 8 presents a comparison of the model's performance when pretrained on the ACDC datasets versus the ImageNet dataset. The results highlight that our method, which utilizes ACDC datasets for pretraining, significantly outperforms the model pretrained on ImageNet across all evaluation metrics. This underlines the value of domain-specific pretraining, as models trained on datasets with characteristics similar to the target domain are more adept at learning relevant features, leading to improved segmentation accuracy.

Several factors contribute to the superior performance of our approach in comparison to

previous methods on the MyoPS dataset. First, our method incorporates a hybrid loss function that combines weighted Dice loss and cross-entropy loss, effectively addressing class imbalance by giving more focus to underrepresented classes. The model's architecture, which integrates attention mechanisms and transformer-based components, further enhances its ability to capture intricate contextual information. Additionally, the use of extensive data augmentation techniques improves the model's ability to generalize, while precise hyperparameter tuning ensures stable and efficient learning. Collectively, these factors enable our approach to outperform existing state-of-the-art models, as demonstrated by the higher DSC scores in Table 5.7.

5.6 Conclusion

In this chapter, we introduced the two-stage transfer learning (2-TLViT) approach, which significantly improves the segmentation accuracy of myocardial infarction in CMR images. By leveraging the strengths of ViTs and the TransUNet architecture, along with an extensive pretraining process on both classification and segmentation datasets, 2-TLViT outperforms existing state-of-the-art models in terms of segmentation precision.

The comprehensive evaluation of the proposed model demonstrated notable improvements across various cardiac regions, reinforcing the potential of 2-TLViT as a reliable tool for clinical applications in cardiovascular disease diagnosis. The method's ability to improve segmentation accuracy, particularly in challenging regions such as infarcted myocardium, is a testament to the model's robustness and adaptability to real-world medical data.

The promising results achieved by 2-TLViT highlight its potential to advance the field of CAD for cardiovascular imaging. By optimizing transfer learning strategies and incorporating robust pretraining techniques, 2-TLViT not only improves segmentation performance but also reduces the reliance on large, annotated datasets, making it more feasible for clinical adoption.

In conclusion, 2-TLViT represents a significant step forward in the application of deep learning for medical image analysis, offering a powerful approach to myocardial infarction segmentation that could have a substantial impact on the early detection and treatment of cardiovascular diseases. Future work should focus on further optimizing the model's performance across diverse patient populations and imaging modalities, as well as exploring its integration into clinical workflows for real-time, reliable diagnosis.

Chapter 6

Conclusions and Perspectives

This thesis has made important contributions to the field of medical imaging, particularly in the application of deep learning for the segmentation of cardiovascular structures. Two novel approaches were developed and thoroughly evaluated: FCTransNet, an ensemble model for segmenting cardiac structures in short-axis CMR scans, and 2-TLViT, a two-stage TL approach for enhancing myocardial infarction segmentation in CMR images. These methods offer significant improvements over existing models, demonstrating their potential for clinical applications in the early detection and diagnosis of cardiovascular diseases.

FCTransNet, by integrating ViTs in an ensemble framework and utilizing image fusion techniques (IWST) and ROI extraction, achieved superior segmentation accuracy in cardiac MRI, particularly for the LV and Myo. However, the approach also faced several limitations. One of the key challenges was the need for high-quality annotated datasets, which are often scarce in the medical field. Additionally, FCTransNet demands substantial computational resources, which can limit its deployment in resource-constrained settings. Furthermore, the generalizability of FCTransNet to other imaging modalities (such as CT scans) or different patient populations has yet to be fully validated. Future research will focus on mitigating these limitations by exploring transfer learning from larger, diverse datasets, employing model compression techniques like pruning and quantization to reduce computational costs, and adapting the model to other medical imaging modalities. Additionally, methods such as semi-supervised learning could be explored to reduce the reliance on labeled data.

Similarly, 2-TLViT addressed the issue of myocardial infarction segmentation by combining the strengths of ViTs and the TransUNet architecture. By leveraging extensive pretraining on both classification and segmentation datasets, this approach achieved superior segmentation performance, surpassing existing state-of-the-art models. Despite

these promising results, 2-TLViT also faced challenges. The approach still requires significant computational resources for pretraining and fine-tuning, and its generalizability to other imaging modalities or diverse patient groups needs further investigation. To address these challenges, future work will explore the use of semi-supervised and unsupervised learning techniques, such as pseudo-labeling and GANs, to reduce dependency on large annotated datasets. Additionally, the model could be extended to other pathologies, such as brain tumor segmentation, and tested on different imaging modalities like CT or PET scans.

Both contributions underscore the importance of ensemble learning, transfer learning, and advanced model architectures in medical imaging. The methods presented in this thesis have shown promise in improving diagnostic accuracy, particularly in the context of cardiovascular diseases, and have the potential to be applied to other domains of medical imaging. However, to fully realize their clinical potential, it will be necessary to address the current limitations related to data availability, computational resources, and generalizability.

In conclusion, the research presented in this thesis advances the field of medical image segmentation, particularly for cardiovascular diseases. The contributions demonstrate that DL models, such as ViTs and TL approaches, can significantly improve segmentation accuracy and hold great potential for clinical implementation. Nevertheless, addressing the limitations of data access, computational efficiency, and model adaptability remains crucial for making these approaches feasible in real-world clinical settings. However, despite the advancements presented, several challenges remain, and further research is required to fully optimize and generalize these techniques for widespread clinical use. Here are some key perspectives we aim to address in future works:

1. **Enhancing Dataset Diversity and Annotation Strategies:** One of the main limitations of this research is the reliance on large, high-quality annotated datasets, which are often difficult to obtain in the medical field. Future research should focus on reducing the dependency on extensive labeled data through techniques like semi-supervised learning, pseudo-labeling, and active learning. Additionally, strategies such as data augmentation and synthetic data generation using Generative Adversarial Networks (GANs) could be explored to create larger and more diverse datasets, thereby improving the model's generalizability across different patient populations and medical imaging modalities.

2. **Improving Computational Efficiency:** The proposed methods, particularly FCTransNet, require substantial computational resources, which can be a significant barrier for real-time clinical applications. Future work should investigate model compression techniques, including pruning, quantization, and knowledge

distillation, to reduce the computational cost without sacrificing performance. Exploring hardware-specific optimizations and the use of lightweight models could help make these techniques more applicable in resource-constrained environments, such as mobile devices or low-cost diagnostic equipment.

3. **Generalizing Across Different Imaging Modalities and Pathologies:** While this thesis focused on cine-MRI for cardiac segmentation, extending the proposed approaches to other imaging modalities, such as CT, PET, or echocardiography, remains a crucial next step. Future work should explore domain adaptation techniques to adapt the models to different imaging protocols, ensuring that the models can generalize across various imaging techniques. Additionally, applying the methods to other cardiovascular pathologies, such as aortic diseases or ischemic heart disease, will be essential for testing the robustness and versatility of the proposed models.

4. **Integration of Multi-Task Learning:** Future work could also explore integrating multiple tasks within a single deep learning framework. For instance, instead of treating segmentation as a standalone task, it could be integrated with other diagnostic tasks, such as disease classification or lesion detection, in a unified model. This multi-task learning approach could enhance the model's ability to learn shared representations and improve its performance on tasks with limited labeled data.

5. **Leveraging longitudinal and temporal data:** In this thesis, the proposed methods were evaluated on static, single-timepoint images. However, medical imaging data often include longitudinal sequences, such as repeated scans over time. Future research should explore the potential of incorporating temporal information into the models, particularly for tracking the progression of diseases like myocardial infarction. Longitudinal data fusion, where multiple scans from the same patient are combined, could help improve segmentation accuracy and provide more dynamic insights into disease progression.

6. **Broader Clinical Applicability and Multi-Modality Fusion:** A major direction for future work will be the extension of the proposed methods to broader clinical applications, particularly the integration of multi-view, multi-temporal, or multi-modality images. For instance, combining MRI data with CT scans or PET images could provide more comprehensive insights into cardiovascular diseases. This multi-modality fusion approach could lead to more generalized models that are robust across different clinical scenarios, enabling more accurate and reliable diagnoses in a variety of healthcare settings.

7. **Deployment in Clinical Environments:** The ultimate goal of this research is to facilitate the clinical adoption of DL-based CAD systems. Future work should

focus on deploying the developed models in real-world clinical environments, including prospective studies and collaborations with hospitals or medical centers. This will help assess the real-world performance of the models and refine them based on clinician feedback. Ensuring that the models are interpretable and can provide actionable insights will be critical for their acceptance in clinical practice.

Bibliography

[1] WHO. World health organization, Aug 2024. URL `https://www.who.int`.

[2] Christopher M Kramer, Jörg Barkhausen, Chiara Bucciarelli-Ducci, Scott D Flamm, Raymond J Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (cmr) protocols: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):17, 2020.

[3] MC Williams, John H Reid, Graham McKillop, NW Weir, EJR van Beek, NG Uren, and DE Newby. Cardiac and coronary ct comprehensive imaging approach in the assessment of coronary heart disease. *Heart*, 97(15):1198–1205, 2011.

[4] Xiaoyu Sun, Yuzhe Yin, Qiwei Yang, and Tianqi Huo. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. *European Journal of Medical Research*, 28(1):242, 2023.

[5] Yan-Ran Wang, Kai Yang, Yi Wen, Pengcheng Wang, Yuepeng Hu, Yongfan Lai, Yufeng Wang, Kankan Zhao, Siyi Tang, Angela Zhang, et al. Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nature Medicine*, pages 1–10, 2024.

[6] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

[7] Sema Atasever, Nuh Azginoglu, Duygu Sinanc Terzi, and Ramazan Terzi. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clinical imaging*, 94:18–41, 2023.

[8] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[9] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126: 106669, 2023.

[10] Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] J. G. Betts, K. A. Young, and J. A. Wise. *Anatomy and Physiology 2e.* OpenStax, 2021.

[12] K. Carter and M. Rutherford. *Building a Medical Terminology Foundation.* eCampus Ontario, 2020.

[13] Ergashov Behro'zjon Komilovich. Coronary artery disease. *EUROPEAN JOURNAL OF MODERN MEDICINE AND PRACTICE*, 3(12):81–87, 2023.

[14] Daniel M Musher, Michael S Abers, and Vicente F Corrales-Medina. Acute infection and myocardial infarction. *New England Journal of Medicine*, 380(2):171–176, 2019.

[15] Moussa Saleh and John A Ambrose. Understanding myocardial infarction. *F1000Research*, 7, 2018.

[16] Martina Perazzolo Marra, João AC Lima, and Sabino Iliceto. Mri in acute myocardial infarction. *European heart journal*, 32(3):284–293, 2011.

[17] Amy Groenewegen, Frans H Rutten, Arend Mosterd, and Arno W Hoes. Epidemiology of heart failure. *European journal of heart failure*, 22(8):1342–1356, 2020.

[18] Gary Tse. Mechanisms of cardiac arrhythmias. *Journal of arrhythmia*, 32(2):75–81, 2016.

[19] Global Cardiovascular Risk Consortium. Global effect of modifiable risk factors on cardiovascular disease and mortality. *New England Journal of Medicine*, 389 (14):1273–1285, 2023.

[20] NHANES. National health and nutrition examination survey homepage, July 2024. URL `https://www.cdc.gov/nchs/nhanes/index.htm`.

[21] Seth S Martin, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Bethany Barone Gibbs, Andrea Z Beaton, Amelia K Boehme, et al. 2024 heart disease and stroke statistics: a report

of us and global data from the american heart association. *Circulation*, 149(8): e347–e913, 2024.

[22] George A Mensah, Valentin Fuster, Christopher JL Murray, Gregory A Roth, Global Burden of Cardiovascular Diseases, and Risks Collaborators. Global burden of cardiovascular diseases and risks, 1990-2022. *Journal of the American College of Cardiology*, 82(25):2350–2473, 2023.

[23] Matthew D Ritchey. Vital signs: state-level variation in nonfatal and fatal cardiovascular events targeted for prevention by million hearts 2022. *MMWR. Morbidity and Mortality Weekly Report*, 67, 2018.

[24] NVSS. Public use data file documentation, July 2024. URL `https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm`.

[25] Ragavendra R Baliga, Kim A Eagle, William F Armstrong, David S Bach, and Eric R Bates. *Practical Cardiology*. Springer, 2008.

[26] N Aerts, D Le Goff, M Odorico, JY Le Reste, P Van Bogaert, L Peremans, G Musinguzi, P Van Royen, and H Bastiaens. Systematic review of international clinical guidelines for the promotion of physical activity for the primary prevention of cardiovascular diseases. *BMC family practice*, 22:1–21, 2021.

[27] Lien Desteghe, Zina Raymaekers, Mark Lutin, Johan Vijgen, Dagmara Dilling-Boer, Pieter Koopman, Joris Schurmans, Philippe Vanduynhoven, Paul Dendale, and Hein Heidbuchel. Performance of handheld electrocardiogram devices to detect atrial fibrillation in a cardiology and geriatric ward setting. *Ep Europace*, 19(1): 29–39, 2017.

[28] Roshan Joy Martis, U Rajendra Acharya, and Hojjat Adeli. Current methods in electrocardiogram characterization. *Computers in biology and medicine*, 48:133–149, 2014.

[29] Martin Geyer, Johannes Wild, Thomas Muenzel, Tommaso Gori, and Philip Wenzel. State of the art—high-sensitivity troponins in acute coronary syndromes. *Cardiology Clinics*, 38(4):471–479, 2020.

[30] Yader Sandoval, Fred S Apple, Simon A Mahler, Richard Body, Paul O Collinson, Allan S Jaffe, International Federation of Clinical Chemistry, and Laboratory Medicine Committee on the Clinical Application of Cardiac Biomarkers. High-sensitivity cardiac troponin and the 2021 aha/acc/ase/chest/saem/scct/scmr guidelines for the evaluation and diagnosis of acute chest pain. *Circulation*, 146 (7):569–581, 2022.

[31] Pavel Poredos, Agata Stanek, Mariella Catalano, and Vinko Boc. Ankle-brachial index: Diagnostic tool of peripheral arterial disease and predictor of cardiovascular risk—an update of current knowledge. *Angiology*, page 00033197241226512, 2024.

[32] Glenn N Levine, Eric R Bates, James C Blankenship, Steven R Bailey, John A Bittl, Bojan Cercek, Charles E Chambers, Stephen G Ellis, Robert A Guyton, Steven M Hollenberg, et al. 2015 acc/aha/scai focused update on primary percutaneous coronary intervention for patients with st-elevation myocardial infarction: an update of the 2011 accf/aha/scai guideline for percutaneous coronary intervention and the 2013 accf/aha guideline for the management of st-elevation myocardial infarction: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the society for cardiovascular angiography and interventions. *Circulation*, 133(11):1135–1147, 2016.

[33] Ziad Issa, John M Miller, et al. *Clinical arrhythmology and electrophysiology: A companion to Braunwald's heart disease*. Elsevier Health Sciences, 2012.

[34] George Mcleod, Kelly Shum, Tripti Gupta, Shourjo Chakravorty, Sergey Kachur, Lisa Bienvenu, Michael White, and Sangeeta B Shah. Echocardiography in congenital heart disease. *Progress in cardiovascular diseases*, 61(5-6):468–475, 2018.

[35] B. Preim and Charl P. Botha. Visual exploration and analysis of perfusion data. In *Visual Computing for Medicine (Second Edition)*, 2014. URL `https://api.semanticscholar.org/CorpusID:78498858`.

[36] Gizeaddis Lamesgin Simegn, Worku Birhanie Gebeyehu, and Mizanu Zelalem Degu. Computer-aided decision support system for diagnosis of heart diseases. *Research Reports in Clinical Cardiology*, pages 39–54, 2022.

[37] Devansh Shah, Samir Patel, and Santosh Kumar Bharti. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6):345, 2020.

[38] Ting-Wei Lin, Po-Yu Huang, and Claire Wan-Chiung Cheng. Computer-aided diagnosis in medical imaging: Review of legal barriers to entry for the commercial systems. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–5. IEEE, 2016.

[39] Yasmina Al Khalil, Sina Amirrajab, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Medical Image Analysis*, 84:102688, 2023.

[40] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database

for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7): 1368–1373, 2018.

[41] Henrique De Melo Ribeiro, Ahran Arnold, James P Howard, Matthew J Shun-Shin, Ying Zhang, Darrel P Francis, Phang B Lim, Zachary Whinnett, and Massoud Zolgharni. Ecg-based real-time arrhythmia monitoring using quantized deep neural networks: A feasibility study. *Computers in Biology and Medicine*, 143:105249, 2022.

[42] Ali Garavand, Ali Behmanesh, Nasim Aslani, Hamidreza Sadeghsalehi, and Mustafa Ghaderzadeh. Towards diagnostic aided systems in coronary artery disease detection: a comprehensive multiview survey of the state of the art. *International Journal of Intelligent Systems*, 2023(1):6442756, 2023.

[43] Kumar G Dinesh, K Arumugaraj, Kumar D Santhosh, and V Mareeswari. Prediction of cardiovascular disease using machine learning algorithms. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–7. IEEE, 2018.

[44] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[45] Javad Hassannataj Joloudari, Sanaz Mojrian, Issa Nodehi, Amir Mashmool, Zeynab Kiani Zadegan, Sahar Khanjani Shirkharkolaie, Roohallah Alizadehsani, Tahereh Tamadon, Samiyeh Khosravi, Mitra Akbari Kohnehshari, et al. Application of artificial intelligence techniques for automated detection of myocardial infarction: a review. *Physiological Measurement*, 43(8):08TR01, 2022.

[46] Brijesh Patel and Partho Sengupta. Machine learning for predicting cardiac events: what does the future hold? *Expert review of cardiovascular therapy*, 18(2):77–84, 2020.

[47] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[48] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[49] Md Farhadul Islam, Sarah Zabeen, Md Humaion Kabir Mehedi, Shadab Iqbal, and Annajiat Alim Rasel. Monte carlo dropout for uncertainty analysis and ecg

trace image classification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 173–182. Springer, 2022.

[50] Ali Haider Khan, Muzammil Hussain, and Muhammad Kamran Malik. Ecg images dataset of cardiac and covid-19 patients. *Data in Brief*, 34:106762, 2021.

[51] Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, Aurelien Bustin, Gastao Cruz, Claudia Prieto, Daniel Rueckert, Julia A Schnabel, and Andrew P King. Deep learning using k-space based data augmentation for automated cardiac mr motion artefact detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 250–258. Springer, 2018.

[52] Steffen E Petersen, Paul M Matthews, Jane M Francis, Matthew D Robson, Filip Zemrak, Redha Boubertakh, Alistair A Young, Sarah Hudson, Peter Weale, Steve Garratt, et al. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of cardiovascular magnetic resonance*, 18(1):8, 2016.

[53] Majd Zreik, Robbert W Van Hamersvelt, Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. A recurrent cnn for automatic detection and classification of coronary artery plaque and stenosis in coronary ct angiography. *IEEE transactions on medical imaging*, 38(7):1588–1598, 2018.

[54] Yiman Liu, Xiaoxiang Han, Tongtong Liang, Bin Dong, Jiajun Yuan, Menghan Hu, Qiaohong Liu, Jiangang Chen, Qingli Li, and Yuqi Zhang. Edmae: An efficient decoupled masked autoencoder for standard view identification in pediatric echocardiography. *Biomedical Signal Processing and Control*, 86:105280, 2023.

[55] Yuhan Ding, Weifang Xie, Kelvin KL Wong, and Zhifang Liao. De-mri myocardial fibrosis segmentation and classification model based on multi-scale self-supervision and transformer. *Computer Methods and Programs in Biomedicine*, 226:107049, 2022.

[56] Alain Lalande, Zhihao Chen, Thomas Decourselle, Abdul Qayyum, Thibaut Pommier, Luc Lorgis, Ezequiel de la Rosa, Alexandre Cochet, Yves Cottin, Dominique Ginhac, et al. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. *Data*, 5(4):89, 2020.

[57] Fumin Guo, Matthew Ng, Maged Goubran, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Graham Wright. Improving cardiac mri convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach. *Medical image analysis*, 61:101636, 2020.

[58] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[59] Chengjin Yu, Yuanting Yan, Shu Zhao, and Yanping Zhang. Pyramid feature adaptation for semi-supervised cardiac bi-ventricle segmentation. *Computerized Medical Imaging and Graphics*, 81:101697, 2020.

[60] Wufeng Xue, Gary Brahm, Sachin Pandey, Stephanie Leung, and Shuo Li. Full left ventricle quantification via deep multitask relationships learning. *Medical image analysis*, 43:54–65, 2018.

[61] Junlin Xian, Xiang Li, Dandan Tu, Senhua Zhu, Changzheng Zhang, Xiaowu Liu, Xin Li, and Xin Yang. Unsupervised cross-modality adaptation via dual structural-oriented guidance for 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(6):1774–1785, 2023.

[62] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

[63] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

[64] Arunava Chakravarty and Jayanthi Sivaswamy. Race-net: a recurrent neural network for biomedical image segmentation. *IEEE journal of biomedical and health informatics*, 23(3):1151–1162, 2018.

[65] Catalina Tobon-Gomez, Arjan J Geers, Jochen Peters, Jürgen Weese, Karen Pinto, Rashed Karim, Mohammed Ammar, Abdelaziz Daoudi, Jan Margeta, Zulma Sandoval, et al. Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE transactions on medical imaging*, 34(7):1460–1473, 2015.

[66] Danielle F Pace, Adrian V Dalca, Tom Brosch, Tal Geva, Andrew J Powell, Jürgen Weese, Mehdi H Moghari, and Polina Golland. Learned iterative segmentation of highly variable anatomy from limited data: Applications to whole heart segmentation for congenital heart disease. *Medical image analysis*, 80:102469, 2022.

[67] Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 80–88. Springer, 2015.

[68] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical image analysis*, 51:21–45, 2019.

[69] Chang Yuwen, Lei Jiang, and Hengfei Cui. Multiple gans guided by self-attention mechanism for automatic cardiac image segmentation. In *Thirteenth International Conference on Graphics and Image Processing (ICGIP 2021)*, volume 12083, pages 509–515. SPIE, 2022.

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

[71] Yongqing Kou, Rongjun Ge, and Daoqiang Zhang. 3d mri cardiac segmentation under respiratory motion artifacts. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 457–465. Springer, 2022.

[72] Shuo Wang, Chen Qin, Chengyan Wang, Kang Wang, Haoran Wang, Chen Chen, Cheng Ouyang, Xutong Kuang, Chengliang Dai, Yuanhan Mo, et al. The extreme cardiac mri analysis challenge under respiratory motion (cmrxmotion). *arXiv preprint arXiv:2210.06385*, 2022.

[73] Hengfei Cui, Chang Yuwen, Lei Jiang, Yong Xia, and Yanning Zhang. Bidirectional cross-modality unsupervised domain adaptation using generative adversarial networks for cardiac image segmentation. *Computers in Biology and Medicine*, 136: 104726, 2021.

[74] Colin Decourt and Luc Duong. Semi-supervised generative adversarial networks for the segmentation of the left ventricle in pediatric mri. *Computers in Biology and Medicine*, 123:103884, 2020.

[75] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023.

[76] Zhenyin Fu, Jin Zhang, Ruyi Luo, Yutong Sun, Dongdong Deng, and Ling Xia. Tf-unet: An automatic cardiac mri image segmentation method. *Math. Biosci. Eng*, 19(5):5207–5222, 2022.

[77] Fatemeh Taheri Dezaki, Zhibin Liao, Christina Luong, Hany Girgis, Neeraj Dhungel, Amir H Abdi, Delaram Behnami, Ken Gin, Robert Rohling, Purang Abolmaesumi, et al. Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss. *IEEE transactions on medical imaging*, 38(8):1821–1832, 2018.

[78] Hadrien Reynaud, Athanasios Vlontzos, Benjamin Hou, Arian Beqiri, Paul Leeson, and Bernhard Kainz. Ultrasound video transformers for cardiac ejection fraction estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 495–505. Springer, 2021.

[79] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[80] Eric Z Chen, Xiao Chen, Jingyuan Lyu, Yuan Zheng, Terrence Chen, Jian Xu, and Shanhui Sun. Real-time cardiac cine mri with residual convolutional recurrent neural network. *arXiv preprint arXiv:2008.05044*, 2020.

[81] AVP Sarvari and K Sridevi. An optimized ebrsa-bi lstm model for highly undersampled rapid ct image reconstruction. *Biomedical Signal Processing and Control*, 83:104637, 2023.

[82] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19 (2020). *arXiv preprint arXiv:2003.13865*, 230, 2003.

[83] Ming Zhao, Xinhong Liu, Hui Liu, and Kelvin KL Wong. Super-resolution of cardiac magnetic resonance images using laplacian pyramid based on generative adversarial networks. *Computerized Medical Imaging and Graphics*, 80:101698, 2020.

[84] Youssef Skandarani, Nathan Painchaud, Pierre-Marc Jodoin, and Alain Lalande. On the effectiveness of gan generated cardiac mris for segmentation. *arXiv preprint arXiv:2005.09026*, 2020.

[85] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, Alexander J Dick, and Graham A Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal*, 2009.

[86] Cristiana Tiago, Andrew Gilbert, Ahmed Salem Beela, Svein Arne Aase, Sten Roar Snare, Jurica Šprem, and Kristin McLeod. A data augmentation pipeline to generate synthetic labeled datasets of 3d echocardiography images using a gan. *IEEE Access*, 10:98803–98815, 2022.

[87] Jun Lyu, Guangyuan Li, Chengyan Wang, Chen Qin, Shuo Wang, Qi Dou, and Jing Qin. Region-focused multi-view transformer-based generative adversarial network for cardiac cine mri reconstruction. *Medical Image Analysis*, 85:102760, 2023.

[88] Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pages 388–400. Springer, 2023.

[89] Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Ivanov Alexander, Jana Taron, Chintan Parmar, Raza M Alvi, Dahlia Banerji, Mio Uno, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nature communications*, 12(1):715, 2021.

[90] Connie W Tsao and Ramachandran S Vasan. Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, 44(6):1800–1813, 2015.

[91] Michiel J Bom, Petrus M van der Zee, and Jan H Cornel. Anatomical versus functional testing for coronary artery. *N Engl J Med*, 372:1291–300, 2015.

[92] Udo Hoffmann, Quynh A Truong, David A Schoenfeld, Eric T Chou, Pamela K Woodard, John T Nagurney, J Hector Pope, Thomas H Hauser, Charles S White, Scott G Weiner, et al. Coronary ct angiography versus standard evaluation in acute chest pain. *New England Journal of Medicine*, 367(4):299–308, 2012.

[93] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[94] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[95] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz,

and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[96] Qi Chang, Zhennan Yan, Meng Ye, Kanski Mikael, Subhi Al'Aref, Leon Axel, and Dimitris N Metaxas. An unsupervised 3d recurrent neural network for slice misalignment correction in cardiac mr imaging. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 141–150. Springer, 2021.

[97] Ghalib A Bello, Timothy JW Dawes, Jinming Duan, Carlo Biffi, Antonio De Marvao, Luke SGE Howard, J Simon R Gibbs, Martin R Wilkins, Stuart A Cook, Daniel Rueckert, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nature machine intelligence*, 1(2):95–104, 2019.

[98] Marcel Beetz, Julius Ossenberg-Engels, Abhirup Banerjee, and Vicente Grau. Predicting 3d cardiac deformations with point cloud autoencoders. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 219–228. Springer, 2021.

[99] KB De Raad, Karin A van Garderen, Marion Smits, Sebastian R van der Voort, Fatih Incekara, EHG Oei, Jukka Hirvasniemi, Stefan Klein, and Martijn PA Starmans. The effect of preprocessing on convolutional neural networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 655–658. IEEE, 2021.

[100] P Vasuki, J Kanimozhi, and M Balkis Devi. A survey on image preprocessing techniques for diverse fields of medical imagery. In *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, pages 1–6. IEEE, 2017.

[101] Sameera V Mohd Sagheer and Sudhish N George. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61:102036, 2020.

[102] Nema Salem, Hebatullah Malik, and Asmaa Shams. Medical image enhancement based on histogram algorithms. *Procedia Computer Science*, 163:300–311, 2019.

[103] Pengling Ren, Yi He, Yi Zhu, Tingting Zhang, Jiaxin Cao, Zhenchang Wang, and Zhenghan Yang. Motion artefact reduction in coronary ct angiography images with a deep learning method. *BMC Medical Imaging*, 22(1):184, 2022.

[104] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):8, 2020.

[105] Jingfeng Lu, Fabien Millioz, François Varray, Jonathan Porée, Jean Provost, Olivier Bernard, Damien Garcia, and Denis Friboulet. Ultrafast cardiac imaging using deep learning for speckle-tracking echocardiography. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2023.

[106] Mingqiang Chen, Lin Fang, Qi Zhuang, and Huafeng Liu. Deep learning assessment of myocardial infarction from mr image sequences. *Ieee Access*, 7:5438–5446, 2019.

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

[108] Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access*, 9:82031–82057, 2021.

[109] Truong Dang, Tien Thanh Nguyen, John McCall, Eyad Elyan, and Carlos Francisco Moreno-García. Two-layer ensemble of deep learning models for medical image segmentation. *Cognitive Computation*, pages 1–20, 2024.

[110] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020.

[111] Boqian Wu, Qiao Xiao, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Decebal Constantin Mocanu, Maurice Van Keulen, and Elena Mocanu. E2enet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. *arXiv preprint arXiv:2312.04727*, 2023.

[112] Yushi Qi, Chunhu Hu, Liling Zuo, Bo Yang, and Youlong Lv. Cardiac magnetic resonance image segmentation method based on multi-scale feature fusion and sequence relationship learning. *Sensors*, 23(2):690, 2023.

[113] Zhaobin Wang, E Wang, and Ying Zhu. Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review*, 53(8):5637–5674, 2020.

[114] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

[115] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE, 2020.

[116] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210:103248, 2021.

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

[118] Annika Reinke, Lena Maier-Hein, and Henning Müller. Common limitations of performance metrics in biomedical image analysis. *Proceedings of the Medical Imaging with Deep Learning (MIDL 2021)*, 2021.

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

[120] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444, 2022.

[121] Ian Goodfellow. Deep learning, 2016.

[122] David H Hubel, Torsten N Wiesel, et al. Receptive fields of single neurones in the cat's striate cortex. *J physiol*, 148(3):574–591, 1959.

[123] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[124] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[125] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[126] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[127] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[128] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[129] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

[130] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pages 1059–1071. PMLR, 2021.

[131] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.

[132] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[133] Tran Minh Quan, David Grant Colburn Hildebrand, and Won-Ki Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *Frontiers in Computer Science*, 3:613981, 2021.

[134] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38 (10):2281–2292, 2019.

[135] Raghav Mehta and Jayanthi Sivaswamy. M-net: A convolutional neural network for deep brain structure segmentation. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 437–440. Ieee, 2017.

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

[137] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4151–4160, 2017.

[138] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*, 2017.

[139] Sihan Wang, Lei Li, and Xiahai Zhuang. Attu-net: attention u-net for brain tumor segmentation. In *International MICCAI brainlesion workshop*, pages 302–311. Springer, 2021.

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

[141] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

[142] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[143] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45 (1):87–110, 2022.

[144] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[145] Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, 91:103000, 2024.

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

[147] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

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

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

[150] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[151] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, pages 14–24. Springer, 2021.

[152] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*, 2021.

[153] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6202–6212, 2023.

[154] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[155] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 5233–5242, 2020.

[156] Cian M Scannell, Amedeo Chiribiri, and Mitko Veta. Domain-adversarial learning for multi-centre, multi-vendor, and multi-disease cardiac mr image segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC*

Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, pages 228–237. Springer, 2021.

[157] Xu Zhang, Zhikui Chen, Jing Gao, Wei Huang, Peng Li, and Jianing Zhang. A two-stage deep transfer learning model and its application for medical image processing in traditional chinese medicine. *Knowledge-based systems*, 239:108060, 2022.

[158] Jiana Meng, Zhiyong Tan, Yuhai Yu, Pengjie Wang, and Shuang Liu. Tl-med: A two-stage transfer learning recognition model for medical images of covid-19. *biocybernetics and biomedical engineering*, 42(3):842–855, 2022.

[159] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[160] Xingchang Huang, Yanghui Rao, Haoran Xie, Tak-Lam Wong, and Fu Lee Wang. Cross-domain sentiment classification via topic-related tradaboost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[161] Yufei Gao, Yameng Zhang, Hailing Wang, Xiaojuan Guo, and Jiacai Zhang. Decoding behavior tasks from brain activity using deep transfer learning. *Ieee Access*, 7:43222–43232, 2019.

[162] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, page 100216, 2023.

[163] Maria Baldeon Calisto and Susana K Lai-Yuen. Adaen-net: An ensemble of adaptive 2d–3d fully convolutional networks for medical image segmentation. *Neural Networks*, 126:76–94, 2020.

[164] Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023.

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

[166] Ziyang Wang, Jian-Qing Zheng, and Irina Voiculescu. An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers. In *Annual Conference on Medical Image Understanding and Analysis*, pages 494–507. Springer, 2022.

[167] Chunyu Fan, Qi Su, Zhifeng Xiao, Hao Su, Aijie Hou, and Bo Luan. Vit-frd: A vision transformer model for cardiac mri image segmentation based on feature recombination distillation. *IEEE Access*, 2023.

[168] Lassaad Ben Ammar, Karim Gasmi, and Ibtihel Ben Ltaifa. Vit-tb: ensemble learning based vit model for tuberculosis recognition. *Cybernetics and Systems*, 55(3):634–653, 2024.

[169] Jianwei Qiu, Jhimli Mitra, Soumya Ghose, Camille Dumas, Jun Yang, Brion Sarachan, and Marc A Judson. A multichannel ct and radiomics-guided cnn-vit (radct-cnnvit) ensemble network for diagnosis of pulmonary sarcoidosis. *Diagnostics*, 14(10):1049, 2024.

[170] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2022.

[171] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[172] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022.

[173] Christoforos Galazis, Huiyi Wu, Zhuoyu Li, Camille Petri, Anil A Bharath, and Marta Varela. Tempera: Spatial transformer feature pyramid network for cardiac mri segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 268–276. Springer, 2021.

[174] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022.

[175] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's

clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.

[176] Xiaoniu Yang and Xiaolin Tian. Transnunet: Using attention mechanism for whole heart segmentation. In *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 553–556. IEEE, 2022.

[177] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer, 2021.

[178] Kaizhong Deng, Yanda Meng, Dongxu Gao, Joshua Bridge, Yaochun Shen, Gregory Lip, Yitian Zhao, and Yalin Zheng. Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography. In *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2*, pages 63–72. Springer, 2021.

[179] Yixuan Wu, Kuanlun Liao, Jintai Chen, Jinhong Wang, Danny Z Chen, Honghao Gao, and Jian Wu. D-former: A u-shaped dilated transformer for 3d medical image segmentation. *Neural Computing and Applications*, 35(2):1931–1944, 2023.

[180] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International conference on medical imaging with deep learning*, pages 820–833. PMLR, 2022.

[181] Long Gao, Lei Zhang, Chang Liu, and Shandong Wu. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artificial intelligence in medicine*, 108:101935, 2020.

[182] Zhang Chaoyang, Sun Shibao, Hu Wenmao, and Zhao Pengcheng. Fdr-transunet: A novel encoder-decoder architecture with vision transformer for improved medical image segmentation. *Computers in Biology and Medicine*, 169:107858, 2024.

[183] Yanwen Chong, Ningdi Xie, Xin Liu, and Shaoming Pan. P-transunet: an improved parallel network for medical image segmentation. *BMC bioinformatics*, 24 (1):285, 2023.

[184] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen,

Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[185] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[186] Shuying Xu and Hongyan Quan. Ect-nas: Searching efficient cnn-transformers architecture for medical image segmentation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1601–1604. IEEE, 2021.

[187] Yong Chen, Xuesong Lu, and Qinlan Xie. Atformer: Advanced transformer for medical image segmentation. *Biomedical Signal Processing and Control*, 85:105079, 2023.

[188] Jun Li, Nan Chen, Han Zhou, Taotao Lai, Heng Dong, Chunhui Feng, Riqing Chen, Changcai Yang, Fanggang Cai, and Lifang Wei. Mcrformer: Morphological constraint reticular transformer for 3d medical image segmentation. *Expert Systems with Applications*, 232:120877, 2023.

[189] Fabian Isensee, Paul F Jaeger, Peter M Full, Ivo Wolf, Sandy Engelhardt, and Klaus H Maier-Hein. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*, pages 120–129. Springer, 2018.

[190] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*, pages 111–119. Springer, 2018.

[191] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. Convolutional neural network with shape prior applied to cardiac mri segmentation. *IEEE journal of biomedical and health informatics*, 23(3):1119–1128, 2018.

[192] Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE transactions on medical imaging*, 39(11):3703–3713, 2020.

[193] Georgios Simantiris and Georgios Tziritas. Cardiac mri segmentation with a dilated cnn incorporating domain-specific constraints. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1235–1243, 2020.

[194] Italo Francyles Santos da Silva, Aristofanes Correa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass. A cascade approach for automatic segmentation of cardiac structures in short-axis cine-mr images using deep neural networks. *Expert Systems with Applications*, 197:116704, 2022.

[195] Shunjie Dong, Zixuan Pan, Yu Fu, Qianqian Yang, Yuanxue Gao, Tianbai Yu, Yiyu Shi, and Cheng Zhuo. Deu-net 2.0: Enhanced deformable u-net for 3d cardiac cine mri segmentation. *Medical Image Analysis*, 78:102389, 2022.

[196] Kai-Ni Wang, Xin Yang, Juzheng Miao, Lei Li, Jing Yao, Ping Zhou, Wufeng Xue, Guang-Quan Zhou, Xiahai Zhuang, and Dong Ni. Awsnet: An auto-weighted supervision attention network for myocardial scar and edema segmentation in multi-sequence cardiac magnetic resonance images. *Medical Image Analysis*, 77:102362, 2022.

[197] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021.

[198] Guang Yang, Suhuai Luo, and Peter Greer. A novel vision transformer model for skin cancer classification. *Neural Processing Letters*, 55(7):9335–9351, 2023.

[199] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693, 2020.

[200] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, U Raghavendra, Anjan Gudigar, Wai Yee Chan, K Meenakshi, K Swaraja, Pawel Plawiak, and U Rajendra Acharya. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 42(1):79–107, 2022.

[201] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[202] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2): 151–166, 2020.

[203] Xiang Yu, Jian Wang, Qing-Qi Hong, Raja Teku, Shui-Hua Wang, and Yu-Dong Zhang. Transfer learning for medical images analyses: A survey. *Neurocomputing*, 489:230–254, 2022.

[204] Antong Chen, Tian Zhou, Ilknur Icke, Sarayu Parimal, Belma Dogdas, Joseph Forbes, Smita Sampath, Ansuman Bagchi, and Chih-Liang Chin. Transfer learning for the fully automatic segmentation of left ventricle myocardium in porcine cardiac cine mr images. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*, pages 21–31. Springer, 2018.

[205] Élodie Puybareau, Zhou Zhao, Younes Khoudli, Edwin Carlinet, Yongchao Xu, Jérôme Lacotte, and Thierry Géraud. Left atrial segmentation in a few seconds using fully convolutional network and transfer learning. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 339–347. Springer, 2018.

[206] Belén Serrano-Antón, Alberto Otero-Cacho, Diego López-Otero, Brais Díaz-Fernández, María Bastos-Fernández, Vicente Pérez-Muñuzuri, José Ramón González-Juanatey, and Alberto P Muñuzuri. Coronary artery segmentation based on transfer learning and unet architecture on computed tomography coronary angiography images. *IEEE Access*, 2023.

[207] Yanjie Zhu, Ahmed S Fahmy, Chong Duan, Shiro Nakamori, and Reza Nezafat. Automated myocardial t2 and extracellular volume quantification in cardiac mri using transfer learning–based myocardium segmentation. *Radiology: Artificial Intelligence*, 2(1):e190034, 2020.

[208] Markus Johannes Ankenbrand, David Lohr, Wiebke Schlötelburg, Theresa Reiter, Tobias Wech, and Laura Maria Schreiber. Deep learning-based cardiac cine segmentation: Transfer learning application to 7t ultrahigh-field mri. *Magnetic Resonance in Medicine*, 86(4):2179–2191, 2021.

[209] Xiliang Zhu, Zhaoyun Cheng, Sheng Wang, Xianjie Chen, and Guoqing Lu. Coronary angiography image segmentation based on pspnet. *Computer Methods and Programs in Biomedicine*, 200:105897, 2021.

[210] Ping Gong, Wenwen Yu, Qiuwen Sun, Ruohan Zhao, and Junfeng Hu. Unsupervised domain adaptation network with category-centric prototype aligner for biomedical image segmentation. *IEEE access*, 9:36500–36511, 2021.

[211] Sven Koehler, Tarique Hussain, Zach Blair, Tyler Huffaker, Florian Ritzmann, Animesh Tandon, Thomas Pickardt, Samir Sarikouch, Heiner Latus, Gerald Greil, et al. Unsupervised domain adaptation from axial to short-axis multi-slice cardiac mr images by incorporating pretrained task networks. *IEEE transactions on medical imaging*, 40(10):2939–2953, 2021.

[212] Stéphane Cuenat and Raphaël Couturier. Convolutional neural network (cnn) vs vision transformer (vit) for digital holography. In *2022 2nd International Conference on Computer, Control and Robotics (ICCCR)*, pages 235–240. IEEE, 2022.

[213] Lei Li, Fuping Wu, Sihan Wang, Xinzhe Luo, Carlos Martín-Isla, Shuwei Zhai, Jianpeng Zhang, Yanfei Liu, Zhen Zhang, Markus J Ankenbrand, et al. Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *Medical Image Analysis*, 87:102808, 2023.

[214] Abdul Qayyum, Imran Razzak, Moona Mazher, Xuequan Lu, and Steven A Niederer. Unsupervised unpaired multiple fusion adaptation aided with self-attention generative adversarial network for scar tissues segmentation framework. *Information Fusion*, 106:102226, 2024.

[215] Shuwei Zhai, Ran Gu, Wenhui Lei, and Guotai Wang. Myocardial edema and scar segmentation using a coarse-to-fine framework with weighted ensemble. In *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images: First Challenge, MyoPS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1*, pages 49–59. Springer, 2020.

[216] Junyi Qiu, Lei Li, Sihan Wang, Ke Zhang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr images. *Medical image analysis*, 84:102694, 2023.

[217] Yu Wang, Gang Xiong, Zhen Li, Mingxin Cui, Gaopeng Gou, and Chengshang Hou. Wsnet: A wrapper-based stacking network for multi-scenes classification of dapps. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 163–179. Springer, 2022.