DEEP LEARNING BASED U-NET VARIANTS FOR CARDIAC MRI SEGMENTATION

C. B. Wijesinghe

239375E

Master of Science in Computer Science & Engineering

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2025

DEEP LEARNING BASED U-NET VARIANTS FOR CARDIAC MRI SEGMENTATION

C. B. Wijesinghe

239375E

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science & Engineering

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2025

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 2025/05/21

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science in Computer Science & Engineering Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Dulani Meedeniya

Signature of the Supervisor: Date: 22/ 05/ 2025

DEDICATION

I dedicate this thesis report to the unwavering support and boundless love of my parents, whose encouragement and belief in my abilities have been the foundation of my academic journey. Their sacrifices and dedication to my education have shaped me into the person I am today.

To the University of Moratuwa, my academic home since my undergraduate years, I extend my gratitude for providing an environment that fosters intellectual growth and innovation. The knowledge and skills I have gained here have been instrumental in undertaking this research endeavor.

I express my deepest appreciation to my supervisor, Prof. Dulani Meedeniya, for her invaluable guidance, mentorship, and unwavering support throughout the research process. Her expertise and encouragement have been a guiding light, propelling me forward in my academic endeavors.

I extend heartfelt thanks to the hardworking Cardiologists and Radiologists whose dedication to patient care inspired this research. Their expertise and collaborative spirit have not only elevated the quality of this study, but also emphasized the importance of bridging the gap between medical practitioners and researchers in the pursuit of improved diagnostic tools and methodologies.

I am also grateful to the contributors of publicly available datasets, whose commitment to advancing research by sharing valuable resources has significantly contributed to the success of this research. Their generosity has broadened the scope of this study, enabling a more comprehensive analysis and understanding of cardiac MRI segmentation of ventricular structures and myocardium.

This work is dedicated to all those who have played a role, big or small, in shaping this academic endeavor. Your support and contributions have been instrumental in bringing this thesis to fruition.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Prof. Dulani Meedeniya, whose unwavering support and guidance have been instrumental in the successful completion of this research. Her expertise, encouragement, and valuable insights have significantly contributed to the development and refinement of my research in the field of Cardiac MRI segmentation of ventricular structures and Myocardium.

I would like to express my sincere gratitude to Mr. Dharshana Muthtettugoda for his invaluable support in facilitating access to the High Performance Computing (HPC) server and for his continuous assistance in troubleshooting connectivity issues throughout the course of this research. I also extend my sincere gratitude to the Department of Computer Science and Engineering for granting me access to the HPC resources, which were crucial for the successful completion of my work.

I extend my heartfelt appreciation to the contributors of related studies, whose groundbreaking work has laid the foundation for my research. Their pioneering efforts have enriched my understanding and provided a robust framework for the exploration of cardiac MRI segmentation.

I am also grateful to those who generously made publicly available datasets, without which the empirical validation of my research would not have been possible. Their commitment to advancing scientific knowledge by sharing resources has been a vital aspect of my research journey.

I would like to acknowledge the invaluable feedback and support from my colleagues and peers. Their constructive criticisms, discussions, and encouragement have played a pivotal role in shaping the direction of my research and refining its methodologies.

Last but not least, I want to express my deepest gratitude to my parents for their unwavering support, understanding, and encouragement throughout this academic journey. Their love and encouragement have been a constant source of inspiration, motivating me to strive for excellence.

This thesis represents the culmination of the collective efforts and support from these individuals, and I am truly grateful for their contributions to the successful completion of this research project.

ABSTRACT

Accurate segmentation of ventricular structures and the myocardium from Cardiac Magnetic Resonance (CMR) images is essential for the diagnosis and management of cardiovascular diseases. This study presents a comprehensive approach to cardiac MRI segmentation by developing and evaluating six U-Net variants: Original U-Net, Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer-Based U-Net, each incorporating architectural enhancements tailored to address specific challenges in segmenting complex cardiac anatomy. These architectures incorporate advanced enhancements such as deeper encoder levels, attention mechanisms, residual connections, multi-scale feature fusion, transformer modules, and feedback mechanisms. To improve segmentation robustness, a novel hybrid loss function, combining Dice Loss and Cross-Entropy Loss, was proposed to effectively manage class imbalance and improve segmentation precision. Among the evaluated models, the Feature Pyramid U-Net achieved the highest performance, with Dice coefficients of 0.9388 (Left Ventricle), 0.8759 (Right Ventricle), and 0.8426 (Myocardium), demonstrating its superior ability to capture multi-scale contextual information. To bridge the gap between research and clinical application, an interactive web application was developed and deployed, enabling real-time inference, visual inspection of annotated segmentations, and region-specific descriptions through a user-friendly interface. This work not only advances the design of deep learning architectures for medical image segmentation, but also demonstrates a practical pathway for integrating these models into clinical workflows.

Keywords: Cardiac MRI, Segmentation, U-Net

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
ACDC	Automated Cardiac Diagnosis Challenge
Atn-UN	Attention U-Net
CMR	Cardiac Magnetic Resonance
CMRI	Cardiac Magnetic Resonance Imaging
CNN	Convolutional Neural Network
DC	Dice Coefficient
DL	Deep Learning
ED	End Diastolic
ES	End Systolic
FCN	Fully Convolutional Neural Network
Feed-Res-UN	Feedback Residual U-Net
FP-UN	Feature Pyramid U-Net
FPB	Feature Pyramid Block
FPN	Feature Pyramid Network
GANs	Generative Adversarial Networks
GRU	Gated Recurrent Unit
JC	Jaccard Coefficient
LSTM	Long Short-Term Memory
LV	Left Ventricle
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MYO	Myocardium
O-UN	Original U-Net
Res-UN	Residual U-Net
RNN	Recurrent Neural Network
ROI	Region of Interest
RV	Right Ventricle
SSMs	Statistical Shape Models
w.r.t	with respect to

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Medical image segmentation constitutes a fundamental process in medical imaging, facilitating the extraction of clinically relevant information from complex datasets. The primary objective of medical image segmentation is to partition an image into distinct and semantically meaningful regions, thereby facilitating the identification and delineation of specific anatomical structures or pathological regions. This process is important for numerous medical applications, including diagnosis, treatment planning, and image-guided interventions. Accurate segmentation is particularly challenging in medical images due to inherent variability, noise, and anatomical complexities. Various segmentation techniques, ranging from traditional methods to advanced deep learning approaches, have been employed to address these challenges.

Cardiac Magnetic Resonance (CMR) Imaging has emerged as a powerful and non-invasive modality for assessing the cardiovascular morphology and its function [17]. The segmentation of ventricular structures and myocardium from the CMR images is vital for understanding the cardiac physiology and aiding in the diagnosis and treatment of cardiovascular diseases. This research focuses on the segmentation of Cardiac Magnetic Resonance Imaging (CMRI) scans using a U-Net based architecture, specifically tailored for the segmentation of ventricular structures and myocardium.

1.1 Background

In the recent times, cardiovascular diseases have been recognized as one of the primary contributors to global mortality. In order to effectively diagnose the cardiovascular diseases and reduce the number of deaths, numerous researches have been conducted over the years for the accurate segmentation of the cardiac.

The diagnosis of cardiovascular diseases can be facilitated by examining three significant regions within the anatomical structure of the cardiac, as shown in Figure 1.1. They are the Right Ventricle (RV) , the Left Ventricle (LV) and the Myocardium (MYO) . Extracting various quantitative measures such as RV and LV volume, ejection fraction, wall thickness, and myocardial mass becomes possible by segmenting the cardiac into these anatomically significant regions [13].

Due to the paramount importance of understanding the cardiac physiology, various imaging modalities are being used such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Ultra Sound (US) [18], for the non-invasive evaluation of the cardiovascular functions. Among those techniques, CMRI is considered as the Gold Standard [1, 18], mainly due to possessing an effective spatial resolution which is crucial for assessing structures of small size [19]. Example applications of the above

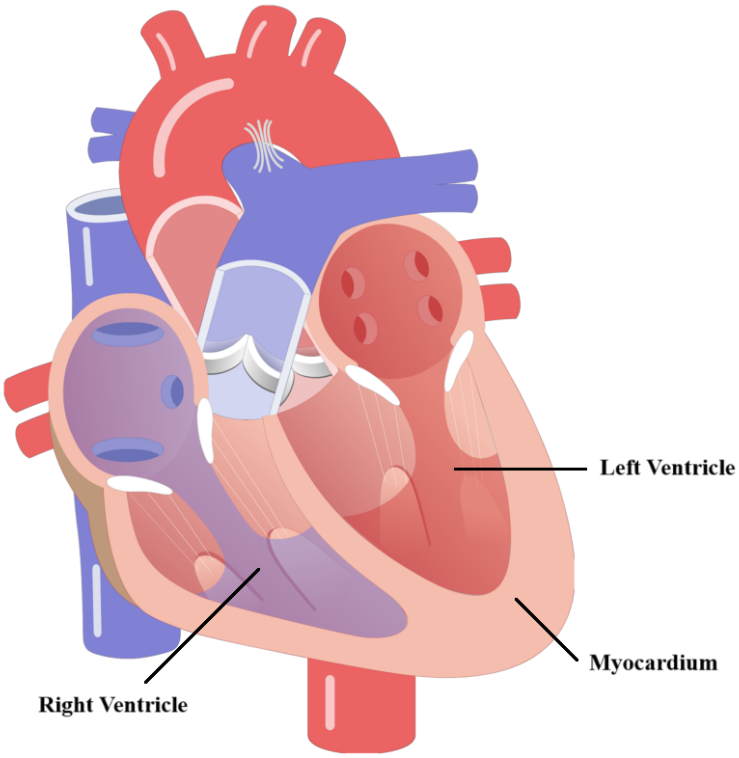

Fig. 1.1: 3D anatomical structure of the cardiac

3 image modalities are depicted in Figure 1.2.

1.2 Research Problem

Accurate segmentation of the Left Ventricle (LV), Right ventricle (RV), and Left Ventricular Myocardium (LVM) in a cardiac short-axis image stack is crucial for the precise computation of quantitative metrics such as myocardial thickness, LV mass, LV and RV Ejection Fractions (EF), and Stroke Volumes (SV). Achieving accurate segmentation during End Diastolic (ED) and End Systolic (ES) phases is therefore essential, involving the precise delineation of the Left Ventricle (LV) endocardium and epicardium, along with the Right Ventricle (RV) endocardium.

The manual labeling process conducted by skilled cardiologists in clinical laboratories is a laborious task. Therefore, in clinical settings, semi-automatic segmentation remains a common practice due to the fully-automatic cardiac segmentation methods having a low accuracy, and the monotonous and time-consuming nature of the manual segmentation process. However, these semi-automated methods are time-consuming, and they are susceptible to variability that can arise both within individual observers (intra-observer variability) and among different observers (inter-observer variability). Therefore, there is a need for a convenient, rapid, reusable, and fully-automated tech-

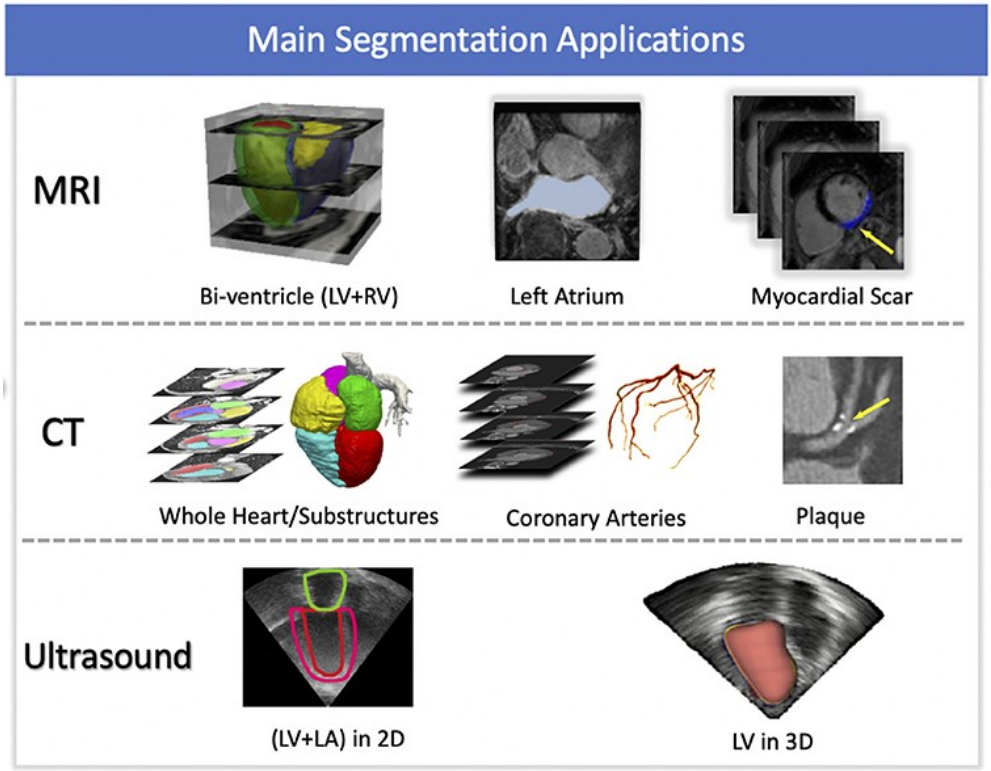

Fig. 1.2: Applications of imaging modalities. Adapted from [1]

nique to segment cardiac regions, facilitating the diagnosis of cardiovascular diseases.

On the other hand, automated segmentation of cardiac MRI scans poses challenges due to below reasons.

- The contrast between the myocardium and surrounding structures is insufficient, characterized by a high contrast between myocardium and blood [18].
- Heterogeneities in brightness within the LV/ RV cavities occur due to variations in blood flow [18].
- The presence of papillary muscles and trabeculae with intensities resembling that of the myocardium [18].
- Non-uniform partial volume effects arise as a result of limited CMR resolution along the long-axis [18].
- Inherent noise is introduced due to motion artifacts and the dynamic nature of the heart [18].
- There is variability in both intensity and shape of heart structures across different pathologies and patients [18].
- The existence of a banding artifact is observed [18].

Despite having these challenges, existing literature suggests the utilization of conventional image segmentation techniques [20] and Deep Learning (DL) techniques [18] for CMRI segmentation. However, the conventional image-processing techniques require either minimal or great user intervention, and the precision of segmentation is significantly influenced by the choice of the training dataset used, which requires significant feature engineering or prior knowledge to obtain better results. On the other hand, DL techniques such as Convolutional Neural Network (CNN) based methods have widely been used for the CMRI segmentation, and have demonstrated acceptable results when dealing with loosely related CMRI slices. However, they tend to encounter challenges such as under/ over segmentation of cardiac structures when the slices are closely connected [21]. Additionally, these conventional image segmentation and DL methods lack the ability to generalize without the need for fine-tuning, when tested on other datasets.

1.3 Motivation and Novelty

Cardiovascular diseases remain a leading cause of mortality worldwide, with accurate diagnosis and treatment playing pivotal roles in improving patient outcomes. Non-invasive imaging techniques, such as Cardiac Magnetic Resonance Imaging (MRI), have emerged as essential tools for assessing cardiac structures and functions. However, the complexity and variability of the heart's anatomy pose significant challenges for automated segmentation, necessitating advanced methodologies.

The primary motivation for this research stems from the need to enhance the precision and efficiency of cardiac MRI segmentation, particularly focusing on ventricular structures and myocardium. Accurate segmentation of these regions is crucial for diagnosing various cardiac conditions, including myocardial infarction, hypertrophic cardiomyopathy, and heart failure. Traditional segmentation methods often fall short in handling the intricate details of cardiac structures, prompting the exploration of deep learning-based approaches.

The novelty of this study lies in the application of U-Net-based variants, a class of convolutional neural networks renowned for their effectiveness in medical image segmentation tasks. By leveraging the inherent capabilities of U-Net architecture, the proposed variants aim to address the challenges associated with cardiac MRI segmentation, providing a more robust and accurate solution. Furthermore, the research explores modifications and improvements to the U-Net model, tailoring it specifically to the intricacies of ventricular structures and myocardial segmentation. The novel adaptations and refinements introduced in this study contribute to advancing the state-of-the-art in cardiac MRI segmentation, offering a promising avenue for improved clinical diagnostics and patient care.

1.4 Research Objectives

The main objectives of this study are as follows.

- Propose U-Net-based architectures for the accurate CMRI segmentation of ventricular structures and myocardium.
- Enhance the accuracy in delineating the boundaries of cardiac structures.
- Compare the performance of the proposed architecture with the existing methods for CMRI segmentation.
- Develop a web application that can be used with real-world clinical settings.

1.5 Research Statement

A robust and accurate system can be developed for the segmentation of ventricular structures and myocardium in Cardiac MRI using U-Net-based architectures, enhancing boundary delineation and improving clinical decision-making. By systematically evaluating multiple U-Net variants within a consistent training pipeline, this study provides a comprehensive comparison of their segmentation performance. Furthermore, the development of a web application ensures the practical applicability of the proposed models in real-world clinical settings, facilitating efficient and reliable cardiac assessment.

1.6 Research Questions

Following research questions are addressed in this study.

- How can accurate segmentation of ventricular structures and myocardium in Cardiac MRI be achieved using U-Net based architectures?
- What techniques can be applied to enhance the accuracy of boundary delineation in cardiac structure segmentation?
- How does the proposed U-Net-based architectures compare with existing methods for Cardiac MRI segmentation in terms of accuracy and efficiency?
- How can the CMRI segmentation model be used for inferencing in clinical settings?

1.7 Thesis Structure

This thesis consists of four main chapters including the Introduction chapter. Chapter 2 will focus on the Literature Review, starting with an overview of CMRI segmentation its importance. It will then review the current clinical practices for CMRI segmentation, followed by a discussion of the available public image datasets and the data pre-processing techniques used for CMRI segmentation studies. The chapter will also explore the various conventional image segmentation techniques and DL-based CMRI segmentation methods. Thresholding, region-growing, pixel classification, deformable methods, atlas-based and Statistical Shape Model (SSM) based methods are discussed in details under the traditional CMRI segmentation methods. Under DL-based methods, CNN, Fully Convolutional Neural Network (FCN), Recurrent Neural Network (RNN), Genarative Adversarial Network (GAN), U-Net and Attention based methods will be discussed in details. A results comparison of the related studies, and widely used evaluation metrics for CMRI segmentation are explained in next two sub-sections. The chapter will conclude by discussing the limitations and challenges in the existing studies.

Chapter 3 presents the detailed Methodology employed for CMRI segmentation using U-Net-based models. It begins by outlining the overall process flow and describing the datasets used in the experiments. The chapter further delves into the data preprocessing techniques applied to enhance model performance. Following this, various U-Net variants, including Original U-Net, Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer-Based U-Net, are discussed in detail. Additionally, the chapter explains the loss function utilized for model optimization, the experimental setup, including training configurations and evaluation protocols, and concludes with a detailed explanation about the web application developed for model inferencing.

Chapter 4 presents the results of the CMRI segmentation experiments. It begins by describing the evaluation metrics employed to assess model performance. Subsequently, the chapter highlights the performance of the models on the ACDC test set, providing insights into their effectiveness. A detailed analysis of segmentation results is provided, showcasing qualitative and quantitative outcomes. Additionally, Dice score and loss graphs are discussed to illustrate the model's learning dynamics and its ability to generalize effectively.

Chapter 5 focuses on discussing the implications of the results and their relevance to the research objectives. It starts by highlighting the contributions of the study, emphasizing how the findings address the proposed research questions. A comparison with existing studies is conducted to situate the research within the broader academic context. The chapter also examines challenges and limitations encountered during the study, such as computational resource constraints, platform dependencies, changes

in external infrastructure, and complexities in model training. Finally, this chapter outlines potential directions for future work to extend and enhance the scope of this research.

Chapter 6 concludes the thesis by summarizing the key findings and their significance. It reflects on how the proposed methodology and results contribute to advancing CMRI segmentation research. The chapter reiterates the study's impact, acknowledges its limitations, and emphasizes future opportunities for exploration and improvement, aiming to inspire further advancements in the field.

CHAPTER 2

LITERATURE REVIEW

Before the Deep Learning (DL) era, traditional image processing and machine learning methods such as model-based, atlas-based and pixel classification approaches were successful in CMRI segmentation; however, those methods often required extensive feature engineering [19]. In contrast, DL algorithms excel in automatically extracting intricate features for object detection and segmentation, by directly learning from the data in an end-to-end fashion [1]. This adaptability allows DL algorithms to be seamlessly applied to diverse image analysis tasks. With the advancements in computer hardware and increased training data availability [13], DL-based segmentation has outperformed traditional methods, particularly evident in the substantial rise of DL-based researches on CMRI segmentation.

This chapter focuses on the Literature Review, beginning with an overview of Cardiac Magnetic Resonance Imaging (CMRI) segmentation and its significance. The chapter proceeds to examine current clinical practices for CMRI segmentation, exploring available public image datasets and the associated data pre-processing techniques in CMRI segmentation studies. Additionally, the chapter delves into a discussion of conventional image segmentation techniques and presents an in-depth exploration of Deep Learning (DL)-based CMRI segmentation methods. The chapter concludes with a comprehensive comparison of results from related studies, discusses widely used evaluation metrics for CMRI segmentation, and provides insights into limitations and challenges inherent in existing research studies.

2.1 Overview of CMRI Image Segmentation

Prior to the rise of Deep Learning (DL), conventional image processing and Machine Learning (ML) techniques, such as atlas-based and model-based approaches, demonstrated commendable performance in CMRI segmentation [1]. Nonetheless, to obtain satisfactory results, these techniques often required substantial feature engineering or prior knowledge, along with minimal or greater user intervention [5, 22].

In contrast to traditional methods, algorithms based on DL demonstrate a superior ability to independently discover features from data. The DL models extracts and leverages meaningful patterns directly from the input data, utilizing an end-to-end approach that eliminates the need for hand-crafted features [23, 24]. This inherent capability of DL-based algorithms could effortlessly be applicable to diverse set of image analysis tasks. With the advent of advanced computer hardware, including Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) [1, 13], coupled with the availability of large volumes of data for training, DL-based segmentation algorithms have

progressively surpassed the performance of earlier state-of-the-art traditional methods, obtaining more popularity in CMRI segmentation researches. This observable shift is shown in Figure 2.1, which depicts a substantial increase in the amount of Deep Learning based papers dedicated to CMRI segmentation in recent years. PubMed [25] search engine was used to query these publications using keywords such as ("cardiac" AND "segmentation") AND ("deep learning" OR "convolutional"). Notably, the volume of publications pertaining to MR image segmentation surpasses that of the other two domains, in all six years.

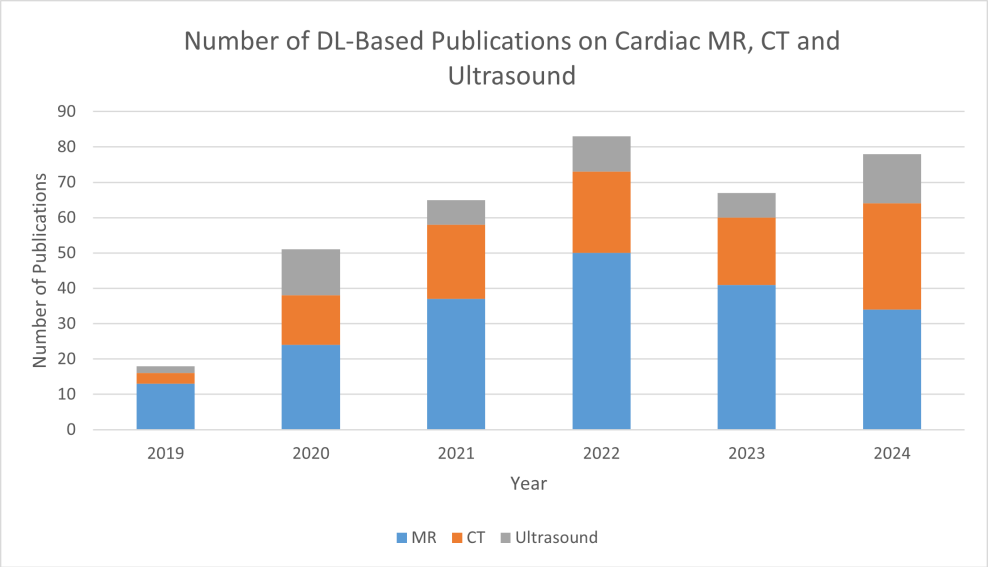

Fig. 2.1: Number of DL-based papers published from January 1, 2019 to December 31, 2024, related to Cardiac MR, CT and US segmentation.

2.2 Current Clinical Practices

Cardiac MRI segmentation is a crucial step in analyzing cardiac images to extract information about the structure and function of the heart. In clinical diagnosis, identification of diastolic and systolic phases, and delineating different regions of interest within the images, such as the LV, RV, myocardium, endocardium, and epicardium is done by medical practitioners [19]. Radiologists and cardiologists play crucial roles in this process, and their interpretation is integral to making accurate diagnoses and treatment decisions.

During cardiovascular MRI, a patient is positioned within the strong magnetic field of a superconducting magnet [26, 27]. Given that heart movement during the cardiac cycle or respiration significantly influences image quality, the electrocardiogram (ECG) is employed to synchronize image acquisition with cardiac-cycle phases, a process known as gating [27, 28]. Typically, images are captured during brief periods of

10 to 20 seconds of breath-holding [28]. To enhance visibility, specialized sequences are applied to manipulate the appearance of blood relative to the myocardium [28], resulting in the creation of static ("dark-blood" or "bright-blood") or dynamic (cine-"bright-blood") images.

These images are typically obtained in various planes, such as short-axis, long-axis, and four-chamber views, to capture different aspects of the heart's anatomy and function [19]. Images may be acquired using different MRI sequences, such as cine imaging for functional assessment, Late Gadolinium Enhancement (LGE) for scar tissue identification, and others [29]. Before the segmentation, pre-processing steps are often applied to enhance image quality and facilitate accurate segmentation.

In clinical settings, below steps are involved in manual segmentation:

- **Image loading and pre-processing:** The MRI images are loaded into a specialized software program such as CMRtools, Heart IT, etc. [29] and pre-processed to remove noise and artifacts.
- **Contouring:** An expert radiologist or cardiologist carefully traces the boundaries of each structure slice by slice, ensuring accuracy and consistency [30].
- **Quality control:** The segmentation results are reviewed and edited if necessary to ensure accurate representation of the cardiac structures [30].
- **Clinical Interpretation:** The final segmented regions, such as the myocardium, endocardium, and epicardium, are used for quantitative analysis. This may involve calculating volumes, ejection fraction, and other functional parameters, depending on the clinical goals [29].
- **Integration with Clinical Workflow:** The segmented results are integrated into the overall clinical workflow, providing valuable information for diagnosis, treatment planning, and monitoring of cardiac conditions.

As highlighted in the aforementioned steps, this process can take hours for a single patient study (i.e. time-consuming and laborious), leading to delays in diagnosis and treatment [13, 21]. Moreover, different experts may segment the same image differently (subjective), leading to inconsistencies in results [21].

Due to the aforementioned challenges, the adoption of automated Cardiac Magnetic Resonance Imaging (CMRI) Segmentation using Deep Learning (DL) is on the rise in clinical practice. This approach offers several advantages, including enhanced accuracy and reproducibility [23]. Additionally, it contributes to increased efficiency by reducing the analysis time and workload for medical practitioners.

2.3 Publicly Available Datasets

Public datasets play a crucial role in advancing research on Cardiac Magnetic Resonance Imaging (CMRI) segmentation. By providing a standardized and diverse set of data, these datasets offer researchers a common ground for developing and evaluating algorithms. This standardized approach contributes to the creation of more accurate and robust models, as researchers can test their methods on a shared benchmark.

Moreover, the use of public datasets promotes fairness and transparency in research practices. It allows the scientific community to validate findings and compare results across different studies, fostering a more open and collaborative environment. Additionally, relying on public datasets enhances the reproducibility of research, as the same dataset is accessible to multiple researchers, enabling them to verify and build upon each other's work.

Table 2.1, adapted from [1], summarizes the publicly available datasets are available for CMRI segmentation studies.

TABLE 2.1: PUBLIC DATASETS FOR CARDIAC MAGNETIC RESONANCE IMAGING (CMRI) SEGMENTATION.

Dataset	**Year**	**# of Subjects**	**Target(s)**	**Related Studies**
York [31]	2008	33	LV, MYO	[32–35]
Sunnybrook [36]	2009	45	LV, MYO	[37–40]
LVSC [41]	2011	200	LV, MYO	[13, 42, 43]
RVSC [44]	2012	48	RV	[45–48]
ACDC [18]	2017	150	LV, MYO, RV	[4, 5, 9, 12, 14, 15, 21, 49, 50]

This study will be mainly conducted using the Automated Cardiac Diagnosis Challenge (ACDC) dataset [18]. The Automated Cardiac Diagnosis Challenge originally took place at the MICCAI 2017 conference, as a two-fold contest: assess the effectiveness of automated methods in accurately segmenting the left ventricular endocardium, epicardium, and the right ventricular endocardium during both ED and ES phases, and evaluate the efficacy of automated methods in classifying examinations into five distinct categories (dilated cardiomyopathy, heart failure with infarction, abnormal RV, hypertrophic cardiomyopathy, and healthy case) [18].

2.4 Data Pre-Processing Techniques

Data pre-processing plays a crucial role in both deep learning and non-deep learning methods for Cardiac MRI segmentation. It enhances the quality of the data, making it

easier for the segmentation algorithms to extract accurate and meaningful information. Below discussed are some commonly used data pre-processing techniques.

- **Image Re-Sampling:** Resizing images to a standard size ensures consistent input for the algorithms [5, 14, 21, 51]. It essentially involves changing the resolution of an image, either up-scaling (increasing) or down-scaling (decreasing) the number of pixels. The choice of re-sampling technique impacts the accuracy and quality of the processed image; hence, it's important to understand the common options available.
  - *Nearest Neighbor:* The value of each pixel in the resized image is assigned to the nearest pixel in the original image [51]. It is computationally efficient, but can lead to blocky artifacts and loss of detail, especially for up-scaling.
  - *Bilinear Interpolation:* This technique uses the weighted average of the four nearest neighbors in the original image to determine the value of each pixel in the resized image [51]. It produces smoother results than nearest neighbor, but can still introduce some blurring, particularly for large up-scaling factors.
  - *Bicubic Interpolation:* This more sophisticated technique uses a polynomial function to estimate the value of each pixel in the resized image based on the surrounding 16 pixels in the original image [10]. It offers a good balance between sharpness and smoothness, making it a popular choice for both up-scaling and down-scaling.

- **Intensity Normalization:** Standardizing the intensity range across images improves segmentation accuracy, especially for deep learning models sensitive to intensity variations. This can involve histogram normalization, scaling, or bias field correction. In DL-based methods, specific transformations like Z-score [22, 51] or min-max scaling, are done to further improve model performance [10].

- **Data Augmentation:** Random transformations such as rotations and translations, and elastic deformations can be applied to augment the dataset and improve model generalization [5, 14, 21, 22, 51]. These methods are widely used in DL-based Cardiac MRI segmentation researches, due to lack of training data available in public datasets to train the models.

- **Region of Interest (ROI) Cropping:** ROI cropping focuses the model on relevant areas, reducing computational complexity. Most of the multi-staged networks proposes a ROI localization network, prior to the main segmentation task [1, 22].

Effective data pre-processing is essential for accurate and reliable CMRI segmentation. By understanding the available techniques and their impact on different algorithms, researchers have optimized their pipelines for specific tasks and datasets, ultimately improving the quality of their segmentation results.

2.5 Traditional CMRI Segmentation Methods

Cardiac MRI segmentation can be tackled through two main approaches: image-driven and model-driven [18, 29]. *Image-driven* approaches leverage techniques such as thresholding, dynamic programming, region growing, pixel classification, graph-cuts and deformable methods, which involve weak prior information [18, 29]. On the other hand, *model-driven* approaches, involving strong prior, leverage powerful Statistical Shape Models (SSMs) and cardiac atlases extracted from labeled training data, capturing average shapes and variations, to guide the segmentation process [18, 29]. In the following subsections, each of these techniques are discussed with their related studies.

2.5.1 Thresholding

Thresholding is a simple tool that uses intensity histograms to analyze the distribution of pixel intensities in CMR image [29]. It looks for prominent modes in the histogram to identify the thresholds. These represent clusters of intensities corresponding to different ROIs. Then, an intensity value (threshold) that separates these modes is chosen for the segmentation [29]. While this approach is straightforward to implement and effective for rapid segmentation, it does have limitations. These include sensitivity to noise and difficulties in accurately delineating ROIs when there are overlaps in intensity. Hence, this technique is often used as an initial step to isolate the ROI [29], before applying any other segmentation technique.

Thresholding is often used with other techniques such as region-growing [29]. The method proposed by Huang *et al.* [52] involves the use of thresholding to differentiate between the blood pool and myocardium. Following this, they apply radial region-growing and utilize convex hulling to identify the boundaries of the endocardium and epicardium. In a similar approach, Lu *et al.* [2] utilize thresholding to convert a ROI into a binary image, facilitating LV localization and the detection of endocardial contours (refer to Figure 2.2). Subsequently, they employ region-growing to achieve the segmentation of the LV epicardium.

2.5.2 Region-Growing

Region-growing is another fundamental image segmentation technique, offering a distinct approach compared to thresholding. Instead of relying solely on intensity values,

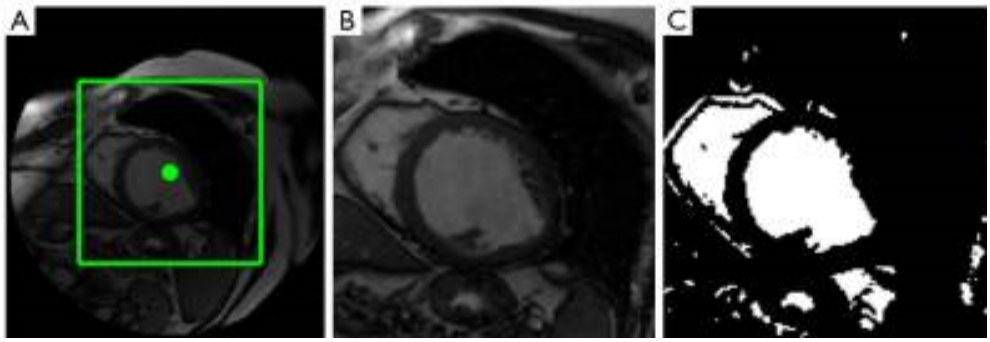

Fig. 2.2: An application of thresholding for LV endocardium segmentation. Adapted from [2]

it leverages spatial connectivity and homogeneity to identify and segment desired regions (such as the myocardium or ventricles) in the CMR images [29].

The region-growing process is started by selecting one or more *seed points* that initiates the growth process, with its neighbors examined for intensity similarity [29]. Neighbors meeting predefined criteria, called growth conditions, are incorporated into the region, becoming new parent pixels for further exploration. This iterative cycle continues until no neighboring pixels satisfy the homogeneity requirements, resulting in the final segmented region [29].

Lee *et al.* [53] and Codella *et al.* [54] both rely on region-growing to identify the complete left ventricle (LV) filled with blood. Their approach starts by automatically pinpointing a seed point within the blood pool. To achieve this, they analyze each pixel inside a window as it moves across the image slices and select the one with the lowest energy as the starting point for region-growing. This ensures that the growth begins within the darkest, most confident area of the blood pool.

2.5.3 Pixel Classification

CMRI segmentation often employs pixel or voxel classification to group individual units in feature space [18, 29]. This feature space can capture information like pixel intensity or texture patterns. Two main approaches exist: unsupervised and supervised.

Unsupervised clustering does not require any manually labeled training data. Techniques like K-means and Expectation-Maximization (EM) automatically group pixels based on similarities in their features. K-means chooses initial cluster centers and assigns pixels to the closest center, then refines the centers based on the assigned pixels [29]. This process iterates until the centers stabilize. EM, on the other hand, statistically models the data, often using Gaussian Mixture Models (GMMs) for cardiac segmentation [29]. Each pixel is assigned to the tissue class that best explains its features.

Supervised classifiers, such as Random Forests and K-Nearest Neighbor (KNN)

leverage manually labeled training data [29]. They learn from these examples to assign classes to new pixels. During training, the classifier adjusts its parameters to minimize misclassification on the training data. Once trained, it can classify new pixels based on what it learned. Annotating training data is, however, laborious and costly, and these methods depend heavily on the quality of the data. Additionally, spatial relationships between pixels are often disregarded, potentially leading to inaccuracies.

Jolly [3] and Hu *et al.* [55] have explored utilizing classification-based methods for cardiac structures. Their techniques rely on Expectation-Maximization (EM) and a Gaussian Mixture Model (GMM) with 3 components, trained on intensity histograms to differentiate tissues. As depicted in Figure 2.3 for Jolly's work [3], this approach can effectively separate the muscle, blood, air, and fat.

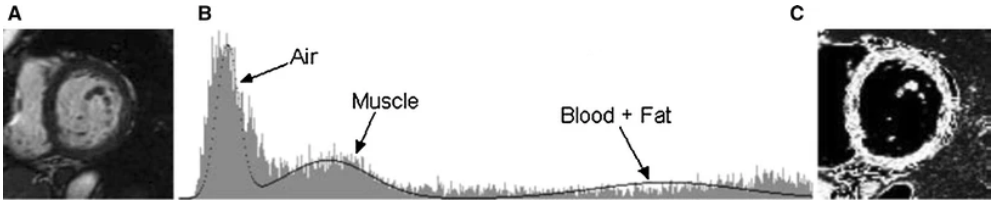

Fig. 2.3: Pixel classification using GMM: **(a)** the input image; **(b)** GMM with 3 components; **(c)** the output image with pixel classification. Adapted from [3]

While KNN classifiers have proven its ability at segmenting cardiac structures like the LV cavity and myocardium against the background, Folkesson *et al.* [56] demonstrated their effectiveness can be further enhanced through feature selection. This method identifies the most influential features for the classifier, enhancing computational efficiency without compromising accuracy. Nonetheless, Bai *et al.* [57] discovered a potentially superior approach for label fusion within multi-atlas cardiac segmentation frameworks: Support Vector Machines (SVMs). Their findings suggest that SVMs may outperform KNN in terms of accuracy, particularly when navigating complex tissue segmentation landscapes, offering a promising avenue for further exploration.

2.5.4 Deformable Methods

In the realm of CMRI segmentation, deformable methods like active contours and level sets offer a distinct approach, guided by explicit models and evolving contours to precisely delineate cardiac structures [18, 19, 29].

Active contours (snakes) iteratively deform to match the object boundaries under the influence of internal smoothness forces and external image-based forces [29]. Key strengths of this method include adaptability to complex shapes, handling of topological changes, and incorporation of prior shape knowledge. However, this method

depends on the initialization of the contour, which may need some level of user interaction [29].

Level sets, an improved method of active contour, represent the evolving contour as an implicit surface within the image domain, governed by a level set function and evolution equation [29]. They excel in handling topological changes, capturing sharp details, and exhibiting reduced sensitivity to initialization.

2.5.5 Atlas-Based Methods

Atlas-based segmentation method, which relies on prior information, uses pre-existing atlases of labeled cardiac structures to guide the segmentation of new images [18, 19, 29]. By borrowing anatomical knowledge from these atlases, atlas-based methods works as follows.

- **Atlas Selection:** An atlas with similar anatomical characteristics to the target image is chosen from a pre-built library [19, 29].
- **Atlas Registration:** The atlas is warped (deformed) to match the target image using registration algorithms [19, 29]. This ensures close alignment of corresponding anatomical landmarks between the two images.
- **Label Transfer:** Once the atlas is aligned, the segmentation labels from the atlas are transferred to the target image [19]. This assigns each pixel in the target image with the corresponding tissue label from the atlas.
- **Refinement and Correction:** The transferred labels might require further refinement or correction, especially in regions with poor registration or significant anatomical differences. This can involve manual adjustments or incorporating other segmentation techniques [19].

Multi-atlas segmentation has made significant strides in tackling complex structures like the RV. In their work, Bai *et al.* [57] leveraged this approach to effectively delineate both the internal and external boundaries of the RV. To optimize cost and accuracy, they employed two key strategies: atlas selection, which identifies the most relevant atlases for a specific image to reduce the search space, and locally-weighted label fusion, which prioritizes atlas labels based on their spatial proximity to the target image, ensuring a contextually-aware segmentation.

2.5.6 Statistical Shape Models (SSM)

In CMRI segmentation, Statistical Shape Models (SSMs) offer a unique approach by leveraging prior information such as statistical information about anatomical shapes and variations, to guide the segmentation [20, 29].

A diverse set of CMR images with manually labeled anatomical structures (e.g. ventricles, myocardium) is collected, and key anatomical landmarks are identified on each image to capture shape outlines [19]. Then, these landmark coordinates are analyzed to extract the mean shape and shape modes. Mean shape refers to the average shape representing the typical structure and shape modes refer to key variations that describe how the shape commonly deviates from the mean [19, 29]. Then these landmark points are aligned across all images using Procrustes analysis, ensuring a common coordinate system. Then, Principal Component Analysis (PCA) is applied to this aligned dataset to identify the most significant modes of shape variation [19]. Then the shape model is constructed using the mean shape and Principal Components (PCs), defining a shape space that encompasses typical anatomical variations.

In CMRI segmentation, two model-based approaches stand out for their frequent use: Active Shape Models (ASMs) and Active Appearance Models (AAMs) [19]. The Active Shape Model (ASM) operates as a local search algorithm utilizing a Point Distribution Model (PDM) [19]. Whereas the ASM specializes in capturing the geometric structure of the data, the Active Appearance Model (AAM) adopts a generative approach, capable of synthesizing realistic representations of the modeled object [19]. This is achieved by incorporating a comprehensive texture model that includes the mean and primary variation modes alongside the shape model [19].

Ordas *et al.* [58] propose a feature vector that remains consistent even under spatial transformations, making it suitable for the ASM framework. Mitchell *et al.* [59] combine the strengths of ASM and AAM in a hybrid approach. They first utilize AAM to fit the object in the image, then leverage the shape information from ASM to escape potential local minimums encountered during tracking. Finally, AAM is reapplied for refined fitting. Similarly, Zhang *et al.* [60] present a combined AAM-ASM model that incorporates temporal features to capture object motion. This combination proves advantageous as it overcomes the limitations of using either ASM or AAM alone.

2.6 Deep Learning Based CMRI Segmentation Methods

DL-based methods have emerged as powerful tools in medical image analysis, particularly in tasks such as cardiac MRI segmentation. CMRI segmentation involves the precise identification and delineation of anatomical structures within the heart, including the LV, RV, myocardium, endocardium, and epicardium [1]. Traditional CMRI segmentation methods often struggled with the complexity and variability of cardiac structures [5, 23], leading to the adoption of DL techniques. These methods leverage Convolutional Neural Networks (CNNs) and other deep architectures to automatically learn hierarchical features from cardiac MRI data, enabling accurate and efficient segmentation.

2.6.1 Convolutional Neural Network (CNN) Based Methods

CNNs are a class of deep learning models that have proven to be highly effective in image analysis tasks, including medical image classification, object localization and detection, and segmentation tasks.

Main components of a CNN include convolutional layers, activation functions, pooling layers and fully connected layers [1, 24]. The architecture diagram of a CNN is shown in Figure 2.4.

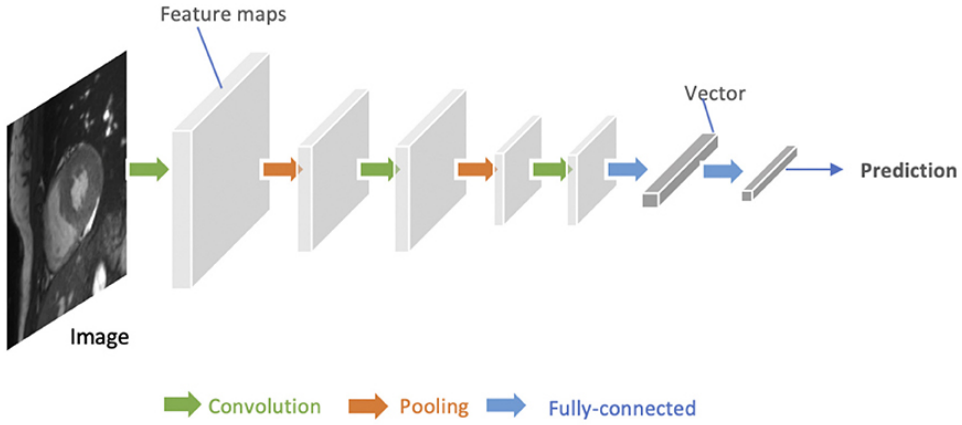

Fig. 2.4: Architecture of a CNN. Adapted from [1]

An explanation of the key components used in a CNN is provided below.

- **Convolutional Layers:** These layers apply convolutional operations to the input image, allowing the network to learn local patterns and features. In cardiac MRI, these patterns could represent edges, textures, or specific structures [1, 24].
- **Activation Functions:** Introduce non-linearities to the network, enabling it to learn complex mappings [1]. Non-linear activation functions such as ReLU and Sigmoid help the network model the intricate relationships between image features, enhancing its ability to accurately segment cardiac structures.
- **Pooling Layers:** Pooling layers downsample the spatial dimensions of the input, reducing computational complexity and focusing on the most essential features. It helps the network maintain translational invariance and reduces sensitivity to small changes in spatial location, making it more robust for cardiac MRI segmentation [1, 24].
- **Fully Connected (Dense) Layers:** These layers establish connections between every neuron in one layer and every neuron in the next layer, capturing global relationships within the data [1]. Fully connected layers at the end of the network map extracted features to the final output, providing a holistic understanding of the input image.

Even though CNNs are widely adopted in image classification tasks, its use can extend to image segmentation applications with minimal adjustments to the network architecture as shown in Figure 2.5. However, this involves dividing each image into patches and training the CNN to predict the class label of the central pixel for each patch [1, 24]. A drawback of this patch-based approach is the need to deploy the network individually for each patch during inference, leading to inefficiency due to redundancy from overlapping patches. In the context of CMRI segmentation, CNNs with fully connected layers are primarily applied for object localization, estimating the bounding box of the target object to reduce computational costs. For more efficient end-to-end pixel-wise segmentation, Fully Convolutional Neural Network (FCN) based approaches are commonly preferred, as discussed in the following subsection.

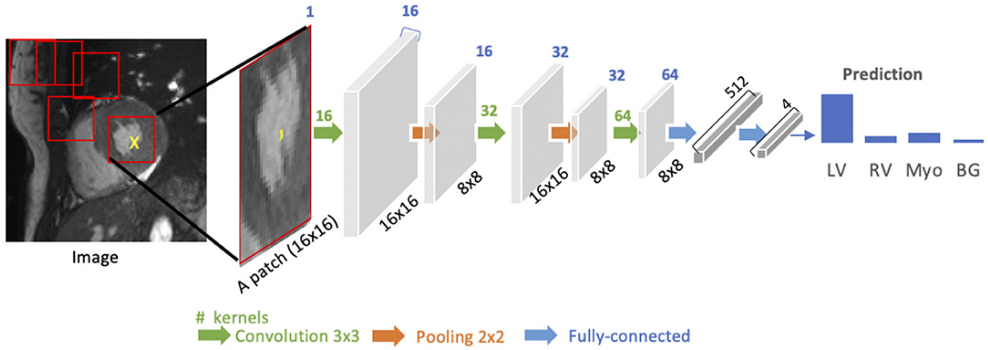

Fig. 2.5: CNN for patch-based segmentation. Adapted from [1]

2.6.2 Fully Convolutional Neural Network (FCN) Based Methods

FCNs are a type of neural network architecture specifically designed for pixel-wise segmentation tasks [61]. Unlike traditional CNNs, FCNs do not have fully connected layers at the end. Instead, they use convolutional layers to directly produce a spatial output, which makes them suitable for tasks such as image segmentation [24].

Figure 2.6 illustrates the encoder-decoder structure of a FCN which enables it to handle input images of arbitrary sizes and generate outputs of the same size. The encoder transforms the input into a high-level feature representation, and the decoder interprets these feature maps, restoring spatial details through a series of upsampling and convolution operations to achieve pixel-wise predictions. Upsampling is accomplished using techniques such as transposed convolutions, unpooling layers, or upsampling layers [1, 24]. Unlike patch-based CNNs, FCNs are trained and applied to entire images, eliminating the need for patch selection.

The simple encoder-decoder structure in Figure 2.6 may have limitations in capturing detailed context information due to potential feature elimination by pooling layers in the encoder [1]. To address this issue, various FCN variants have been proposed,

with the U-Net [6] being a prominent example.

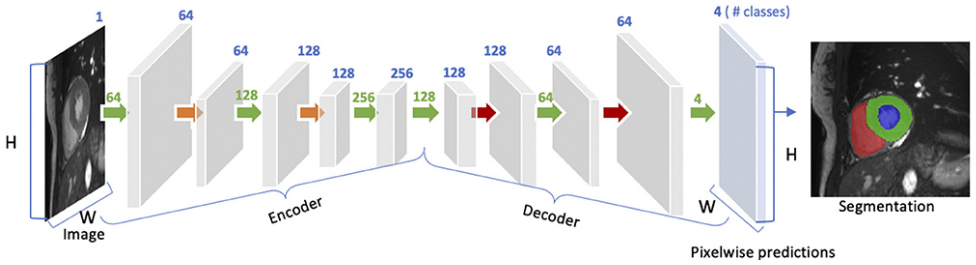

Fig. 2.6: Architecture diagram of a Fully Convolutional Neural Network (FCN). Adapted from [1]

I. F. S. da Silva *et al.* introduced a cascading strategy for the segmentation of CMRI. The initial segmentation of cardiac structures employed a modified FCN, utilizing ROIs extracted from a U-Net during the ROI extraction stage [4]. The FCN architecture, resembling a U-Net, includes contraction and expansion paths along with skip connections, as illustrated in Figure 2.7. In the contraction path, features are extracted using an EfficientNet B3. The expansion path incorporates five convolutional blocks known as Decoder Blocks, each employing the Attention mechanism to generate segmentation masks for the objects of interest.

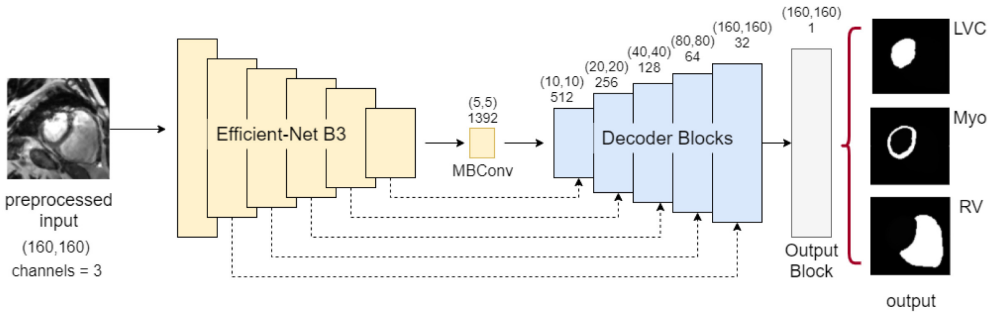

Fig. 2.7: Proposed FCN architecture for initial segmentation. Adapted from [4]

2.6.3 Recurrent Neural Network Based Methods

Recurrent Neural Network (RNN) offer a unique perspective for CMRI segmentation tasks due to its ability to handle sequences by capturing the temporal dynamics of the heart, making it a valuable tool in this specialized domain [1]. Unlike static images, Cardiac MRIs often involve sequences, like cine MRIs capturing the heart's beating. RNNs excel at processing such sequences, utilizing their *memory* to learn dependencies between consecutive frames [5].

Long Short-Term Memory (LSTM) is a common RNN architecture used for effectively capturing the long-range dependencies of the full cardiac cycle [5, 24]. More-

over, Gated Recurrent Unit (GRU) , which is simpler than LSTMs, is another type of RNN architecture used for capturing short-range dependencies.

C. Yutian *et al.* in their study, proposed a Temporal Consistency Network (TCN), which is based on a Residual U-Net [5]. The TCN comprises of hierarchical ConvLSTMs which are bi-directional. The network structure of the TCN is shown in Figure 2.8. In the ACDC dataset [18], the frames in the CMR image exhibit a strong correlation between consecutive frames, leading to the potential propagation of prediction errors from the first frame, particularly due to brightness heterogeneity, to subsequent frames in the CMR sequence. To address this issue, [5] implemented a bi-directional training approach.

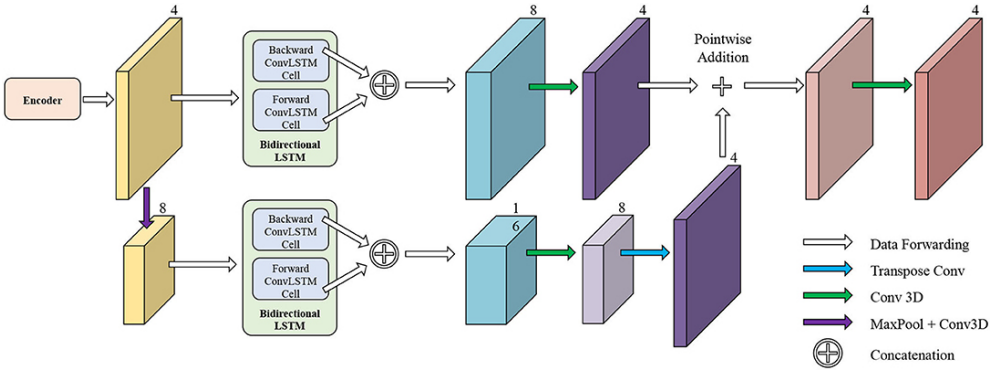

Fig. 2.8: The architecture of TCN. Adapted from [5]

In conclusion, RNNs offer a promising approach for CMRI segmentation, leveraging their unique ability to handle temporal dynamics.

2.6.4 Generative Adversarial Network Based Methods

While CNNs are the workhorses of image segmentation, Generative Adversarial Networks (GANs) offer a unique and exciting approach for Cardiac MRI segmentation. By harnessing the power of synthetic data generation, GANs can address critical challenges and improve segmentation accuracy [1, 62]. A GAN architecture used for CMRI segmentation is shown in Figure 2.9.

While the *Generator* in GANs (i.e. segmentation network in Figure 2.9) aims to generate realistic and anatomically accurate Cardiac MRI images [62], often conditioned on existing segmentation masks or other relevant information, the *Discriminator* acts as a critic, trying to distinguish real images from those generated by the generator [1, 62]. The *Adversarial Training* of GANs refers to a cat-and-mouse game. The generator improves its ability to fool the discriminator by creating more realistic images, while the discriminator becomes sharper at identifying fakes. Over time, both networks become better at their tasks.

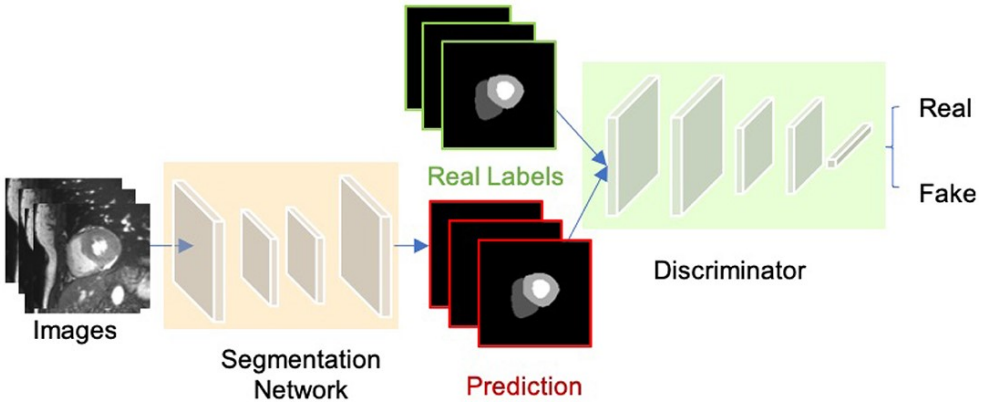

Fig. 2.9: The architecture of GAN. Adapted from [1]

One of the main benefits of GANs for CMRI segmentation is the ability to generate large amounts of realistic synthetic data, addressing the scarcity of high-quality labeled Cardiac MRI data. This can improve model generalizability and robustness to variations in image characteristics [63].

Overall, GANs hold significant promise for revolutionizing Cardiac MRI segmentation. By addressing data scarcity and improved model training, GANs can contribute to more accurate diagnoses and personalized treatment options in cardiology [63–65].

2.6.5 U-Net Based Methods

U-Net and its variants [24, 66–68] are popular CNN architectures widely used in medical image segmentation tasks, including cardiac MRI segmentation. The original U-Net architecture [6] was introduced by Ronneberger *et al.* in 2015, and has since been adapted and extended to address specific challenges in medical image analysis.

The U-Net architecture consists of a contracting path (encoder), a bottleneck, and an expansive path (decoder) [6] as shown in Figure 2.10. In the contracting path, the input image is processed through a series of convolutional layers (3×3 unpadded convolutions), each followed by a rectified linear unit (ReLU) activation [6, 24]. Moreover, 2×2 max-pooling layers are used to downsample the spatial resolution [6]. At the end of the contracting path, the bottleneck contains a set of convolutional layers, providing a high-level representation of the input image [6]. The expansive path, also known as decoder path, involves upsampling the feature maps using transposed convolutions (also known as deconvolutions or fractionally strided convolutions) [6]. Each block in the expansive path consists of convolutional layers, and ReLU activation [6]. At the final layer, 1×1 convolution is performed to produce segmentation masks, typically with softmax activation for multi-class segmentation [6, 24].

One of the key features of U-Net is that it has skip connections from the contracting path, which concatenates the feature maps from that path to help preserve spatial

information, which might loose from the pooling operations [6].

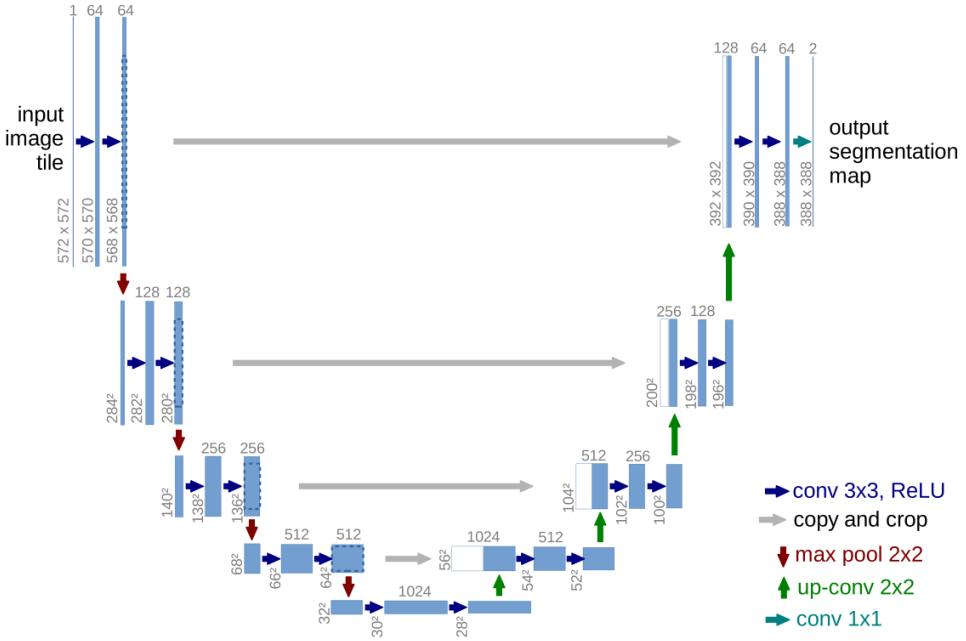

Fig. 2.10: The U-Net architecture. Adapted from [6]

Due to its advantages over the CNN or FCN based methods, various studies have proposed a plethora of variants of U-Net to address specific challenges in CMRI segmentation. Shibuya *et al.* [7] has proposed a Feedback U-Net architecture that uses convolutional LSTM and a softmax cross-entropy loss function and tested its performance on Mouse cell image and Drosophila cell image datasets. Their proposed method utilizes the features acquired during the first round for the second round, and they have used convolutional LSTM layers instead of the convoutional layers [7]. The architecture of the proposed Feedback U-Net is shown in Figure 2.11.

In order to tackle the problems associated with LV segmentation in CMR images, Wu *et al.* [8] proposed a composite model comprises of a CNN and a U-Net. The workflow of the proposed method is depicted in Figure 2.12. The CNN model used in the proposed composite model first locates the ROI, to prevent the inclusion of irrelevant regions with similar gray values in the U-Net [8]. Once the ROI is located, a U-Net model is used to segment the left ventricle [8].

In their study to segment the myocardial in MRI sequences by leveraging temporal information between CMRI sequences, Chen *et al.* [5] proposed an initial segmentation network called *Res U-Net* which is based on the U-Net architecture, as shown in Figure 2.13. It uses a single-channel image as the input and adds a single residual block between each level in the Res U-Net architecture to mitigate the vanishing or exploding gradient problems [5]. In addition to that, instead of concatenating the feature maps from the encoder and decoder, point-wise addition is performed to combine the

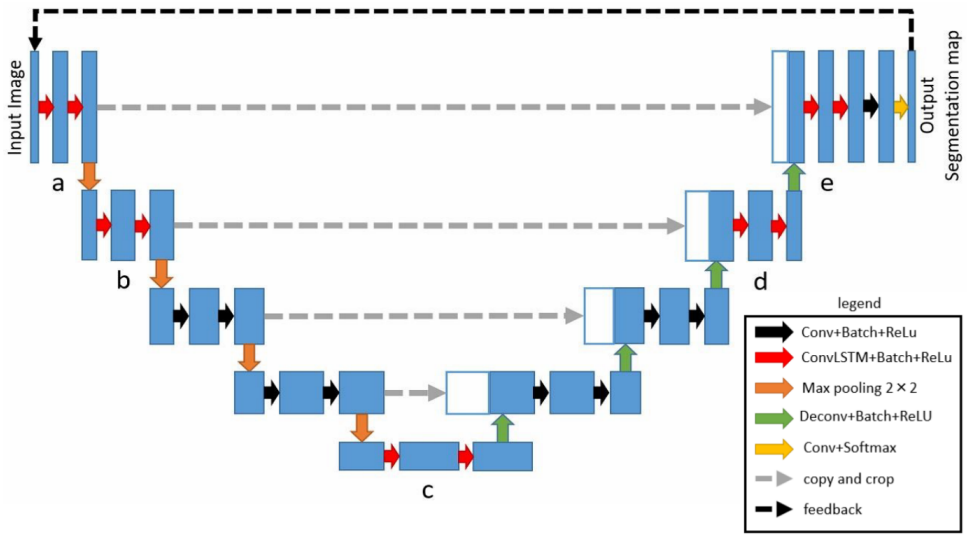

Fig. 2.11: The Feedback U-Net architecture. Adapted from [7]

results of each layer.

A cascaded approach was proposed by da Silva *et al.* [4] for CMRI segmentation, and the U-Net architecture has been used twice in their method. Initially, a U-Net is used for the location and extraction of ROI with the purpose of reduce the scope of the image for processing [4]. Then a FCN is used for the initial segmentation of cardiac

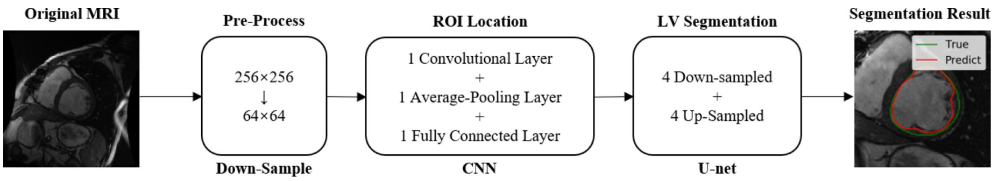

Fig. 2.12: The proposed workflow for CNN + U-Net based composite model. Adapted from [8]

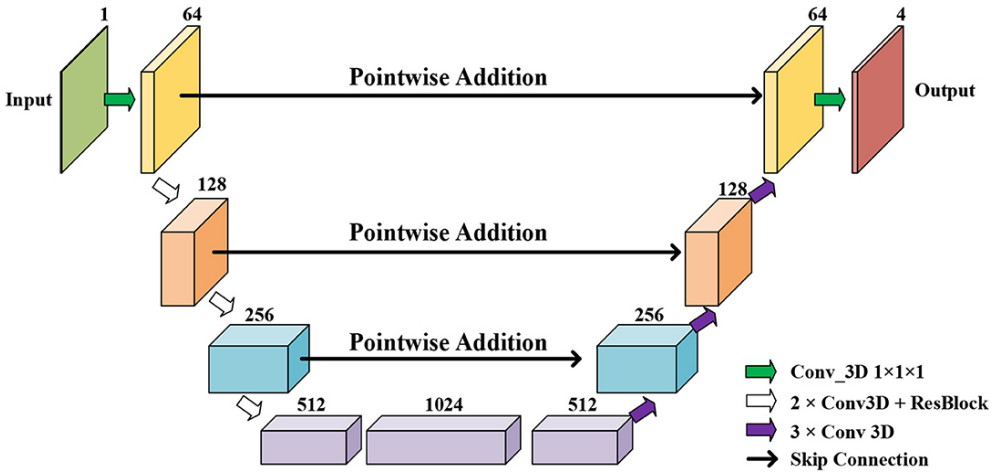

Fig. 2.13: The proposed initial segmentation network: Res U-Net. Adapted from [5]

structures, and finally, a U-Net-based mask reconstruction module is used to further refine the initial segmentation [4].

The public MyoPS 2020 challenge dataset was used by Hengfei *et al.* [51] for the segmentation of abnormal tissues in the Myocardium. Their proposed method comprises of a deep U-Net architecture having 6 layers, as the backbone of the segmentation. Moreover, to explore the performance improvement, 3 additional modules named: Direction Field Module (DFM), Channel self-Attention Module (CAM) and Selective Kernal Module (SKM) have been used, and multiple loss functions and data augmentation methods have been explored [51].

In their seminal work, Ren *et al.* [9] proposed a milti-task learning based U-Net called MTL-UNet, for CMRI segmentation using the ACDC dataset [18]. As shown in Figure 2.14, the model consists of an Edge Extraction (EE) module to extract edge features at various spatial sizes within the encoder path [9]. This addition serves to capture context information in the spatial domain. Furthermore, a fusion-based module is implemented to integrate the extracted edge features obtained from the EE module with both the low and high-level features derived from the original U-Net [9].

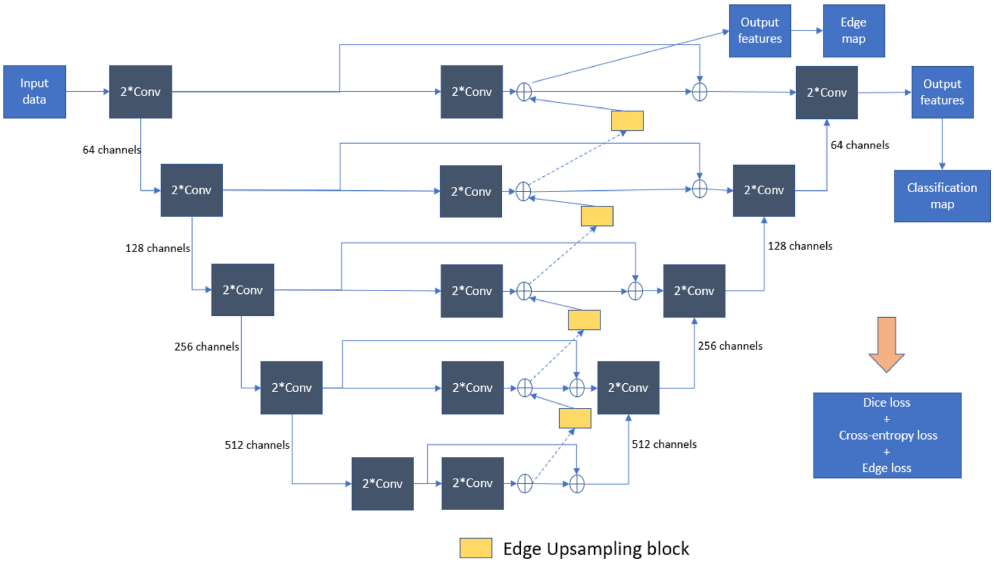

Fig. 2.14: The architecture of the proposed MTL-UNet. Adapted from [9]

Sharan *et al.* [15] proposed an encoder modified U-Net model for the CMRI segmentation of LV, RV and MYO. Different encoders such as DenseNet, ResNet and VGG have been explored in their study, and the U-Net with VGG-based encoder has performed well compared to others [15].

In order to diagnose diseases such as persistent Microvascular Obstruction (MVO) and Myocardial Infarction (MI), different model architectures such as conventional U-Net, U-Net with VGG16, SegNet and Res-U-Net have been explored by Mugahed *et al.* [10] for effective segmentation of MI. An end-to-end AI-based framework (refer

Figure 2.15) is proposed in this study which utilizes the top performing Res U-Net model with CLAHE pre-processing [10].

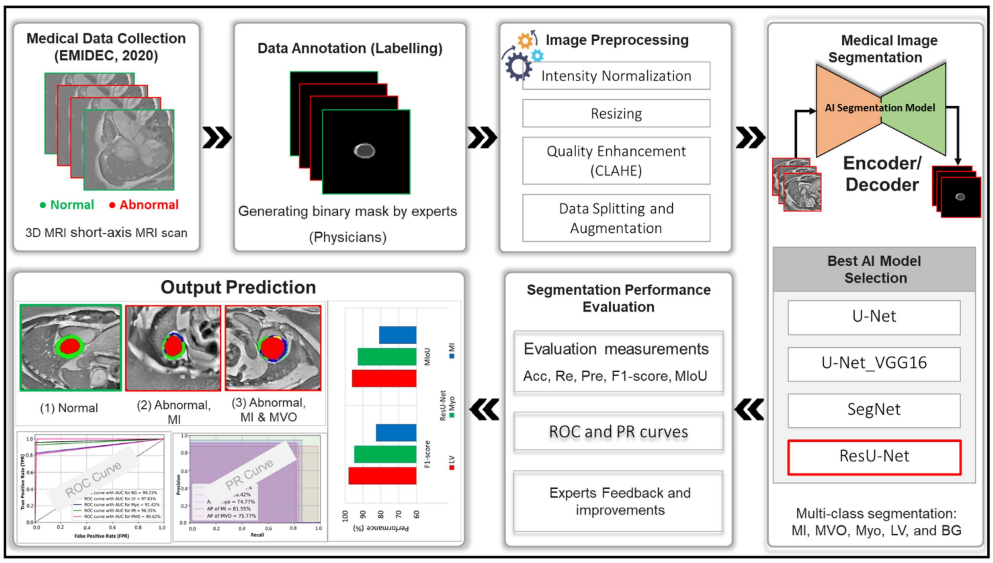

Fig. 2.15: The end-to-end workflow of the AI-based framework. Adapted from [10]

By combining the U-Net architecture with Cross Stage Partial (CSP) method, Chen *et al.* [22] proposed a U-Net-CSP model for CMRI segmentation of LV, RV and aorta structures. The usage of CSP module in the architecture has provided the feature reuse capability and has reduced overfitting [22].

In summary, U-Net and its variants [24] provide powerful tools for cardiac MRI segmentation, enabling accurate and efficient extraction of anatomical structures and abnormalities from medical images. The choice of a specific variant depends on the characteristics of the data and the goals of the segmentation task.

2.6.6 Attention Based Methods

Attention mechanisms in CNNs enable the network to focus on informative regions of the input while suppressing irrelevant ones [11, 24, 69, 70]. This selective attention leads to improved segmentation accuracy, especially for complex structures such as the LV, RV and myocardium.

Oktay *et al.* [11], in 2018, proposed a novel self-attention gating module that can be used with CNN-based models for standard image analysis tasks. They have proposed a grid-based gating mechanism to enhance the specificity of attention coefficients towards local regions [11]. Furthermore, by incorporating a soft-attention technique [24], they have extended its application to medical imaging tasks, integrating it seamlessly with a feed-forward CNN model [11]. As illustrated in Figure 2.16, the authors have expanded upon the conventional U-Net model to introduce the Attention U-Net model, and conducted a performance evaluation on the CT pancreas

segmentation problem [11].

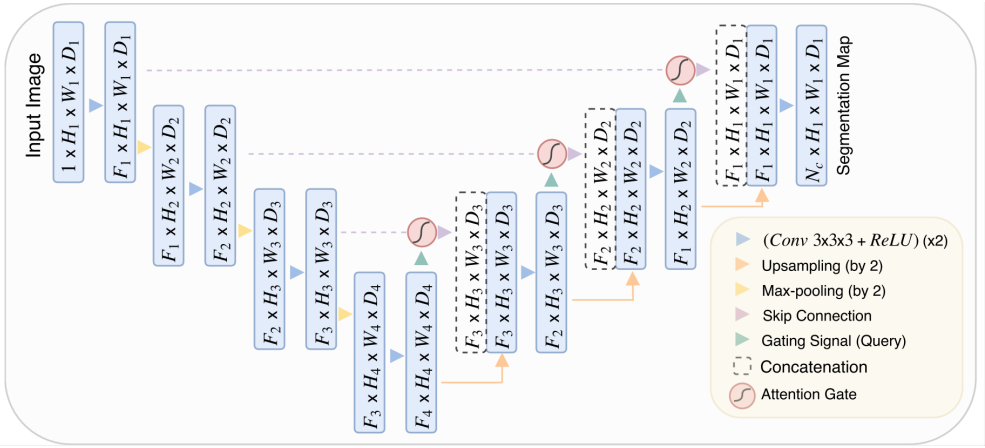

Fig. 2.16: The Attention U-Net model. Adapted from [11]

Robustness of a model and its ability to interpret the results [12] have also been trending research studies associated with CMRI segmentation recently. *Trainable attention* modules, which have trainable parameters, have been used in the models for interpretability in computer vision tasks [12]. In their seminal work, Sun *et al.* [12] proposed a Shape Attentive U-Net (SAUNet) model which comprises of a standard U-Net-based texture stream, and an additional shape stream to capture shape-dependent information (refer Figure 2.17). The output of the secondary shape stream, which produces a shape attentive map, can be used for interpreting the results of the model [12]. The decoder module integrates feature maps from the encoder through skip connections, combining them with the feature maps from lower-resolution decoder blocks that capture additional spatial and contextual information [12]. To enhance interpretability and performance transparency, they have introduced the dual attention decoder block as in Figure 2.17 [12]. This block incorporates two novel components following the standard normalized 3×3 convolution on the concatenated feature maps. These components consist of a channel-wise attention path for enhanced performance, and a spatial attention path for interpretability [12].

In order to create a generalizable DL-based model, Kong *et al.* [62] has utilized the Attention U-Net developed by Oktay *et al.* [11] along with a data augmentation strategy. For the unlabeled data, they have proposed two augmentation methods: create different style of images using a CycleGAN and exchanging low-frequency features among images sourced from various vendors [62]. For the labeled data, spatial augmentation is applied [62].

Hengfei *et al.* [13] have also used the Attention U-Net model [11] with an Input image pyramid and Deep supervised output layers (AID) to propose a multi-scale attention guided U-Net for CMRI segmentation. The proposed architecture is shown in Figure 2.18. By introducing an input image pyramid before each max pooling layer

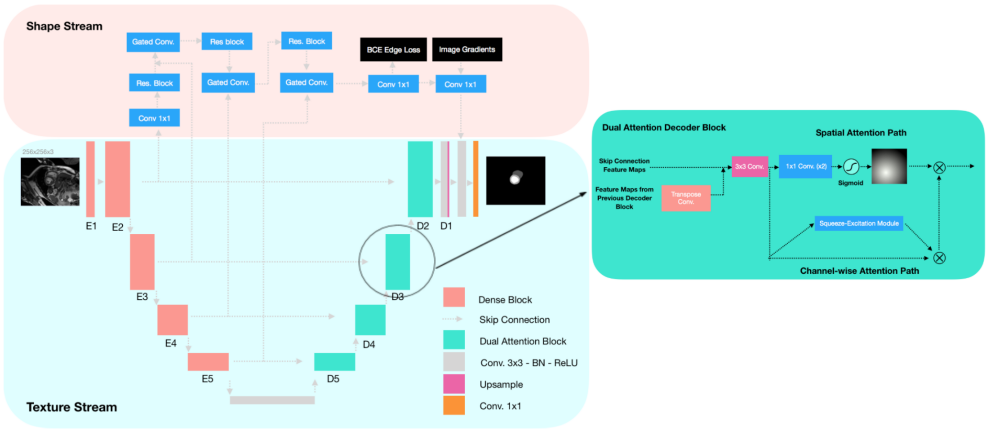

Fig. 2.17: The Shape Attentive U-Net (SAUNet) model. Adapted from [12]

in the encoder, they have addressed the variation in class details and accessibility at different scales [13]. Moreover, it has produced better intermediate feature maps than the conventional U-Net [13].

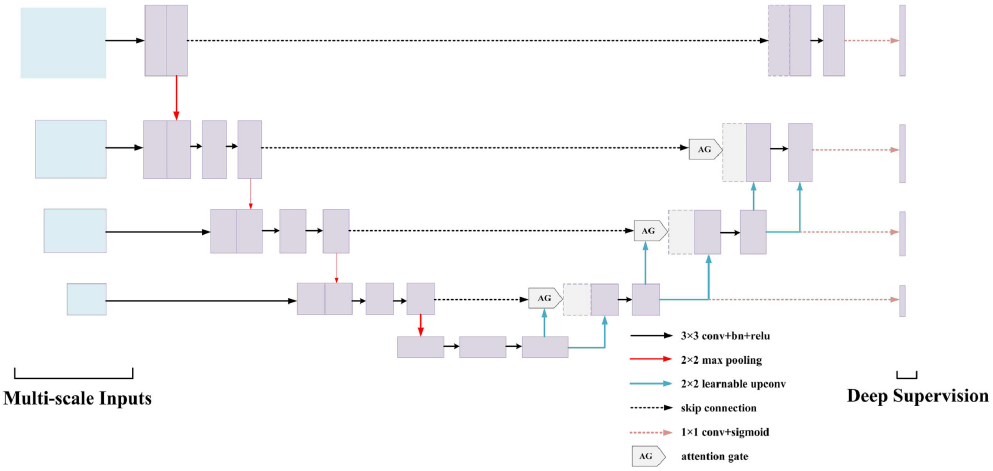

Fig. 2.18: The Attention U-Net model with input image pyramid and deep supervised output layers. Adapted from [13]

For the same objective pursued by Kong *et al.* [62], Kamal *et al.* [14] introduced an attention-guided residual W-Net to improve model generalizability without requiring fine-tuning. As shown in Figure 2.19, the proposed ARW-Net uses multiple attention modules at the decoding path to allow the network to focus on specific components in MR images [14]. Moreover, to effectively address the challenge of vanishing gradients and facilitate feature reuse, ARW-Net adopts a deeply supervised approach by implementing a second path alongside the decoder. Through skip connections, the decoder collects features computed by the encoder [14]. The deep supervision path is constructed by pixel-wise summation of the outputs from each dimension of the feature map in the decoder [14]. This strategy significantly mitigates the vanishing gradi-

ent issue and promotes feature reuse throughout the network. Furthermore, various data augmentation strategies have been employed to enhance the generalizability of the model. As a result, the proposed model has demonstrated improved segmentation performance across multiple datasets without the need for any fine-tuning [14].

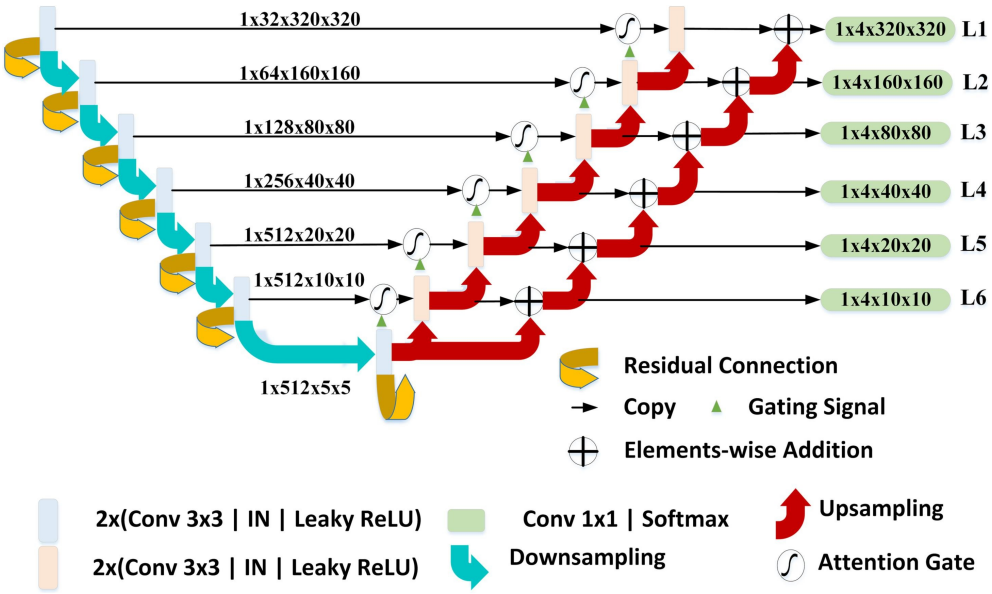

Fig. 2.19: The architecture of the 2D ARW-Net model. Adapted from [14]

2.6.7 State-of-the-Art Model

In their seminal work, Sharan *et al.* [15] introduced two segmentation models for CMRI segmentation based on the ACDC dataset: an encoder-modified Feature Pyramid Network (FPN) and U-Net architectures. The encoder modifications incorporated networks such as VGG, ResNet, and DenseNet. Transfer learning was employed in their study, leveraging pre-trained DenseNet, VGG, and ResNet encoders trained on the ImageNet dataset. This research represents the State-of-the-Art approach for CMRI segmentation using the ACDC dataset.

The first type of model which is the FPN, leveraged a convolutional network's hierarchical feature representation to construct a feature pyramid with high-level semantics across different scales. It processed images of any size while maintaining consistent feature map dimensions. FPN was independent of the backbone architecture and integrated various networks, such as ResNet, DenseNet, InceptionNet, and VGG, as encoders. The architecture of the FPN is shown in Figure 2.20.

FPN consisted of two pathways: bottom-up and top-down. The bottom-up pathway was a feed-forward convolutional network that extracted features at multiple scales with a scaling factor of 2. As it progressed upward, spatial resolution decreased while higher-level features became more abstract. At the top of this pathway, a 1×1 convolu-

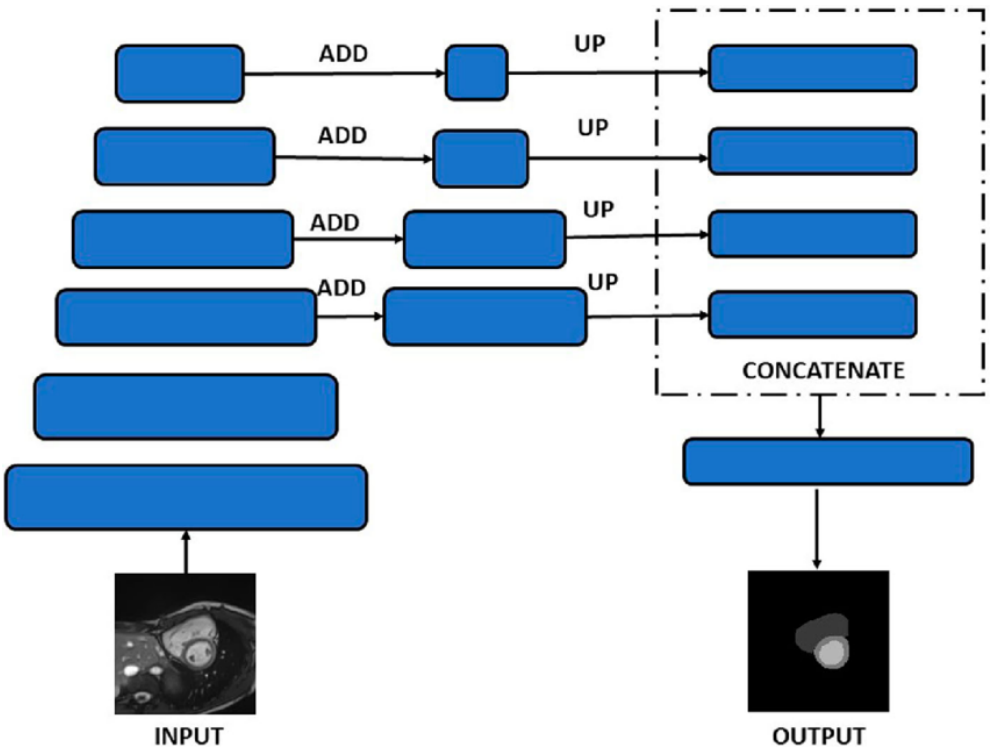

Fig. 2.20: The architecture of the FPN. Adapted from [15]

tion reduced channel depth, followed by two 3×3 convolutions that generated the first segmentation feature map.

In the top-down pathway, feature maps from the bottom-up path were upsampled by a factor of 2 using nearest-neighbor interpolation. A 1×1 convolution was then applied to align feature maps, followed by element-wise addition. Two 3×3 convolutions refined the output before final segmentation. Ultimately, all feature maps (each with 128 channels) were concatenated into a 512-channel representation. A 3×3 convolution with batch normalization and ReLU activation was applied, followed by a final 1×1 convolution to produce the final feature map.

The second type of model which is the U-Net architecture developed by Ronneberger *et al.* [6] was also used with multiple encoders such as VGG, DenseNet and ResNet. The architecture consisted of a contracting path and an expansive path. The contracting path applied two 3×3 convolutions, each followed by a rectified linear unit (ReLU) activation, and a 2×2 max pooling operation with a stride of 2 for downsampling. The number of feature channels doubled after each downsampling step. In the expansive path, upsampling was performed at each stage, followed by a 2×2 convolution that reduced the number of feature channels by half. The upsampled feature maps were then concatenated with the corresponding cropped feature maps from the contracting path. This was followed by two 3×3 convolutional layers with ReLU activation. Cropping was necessary due to the loss of border pixels during convolution.

At the final output layer, a 1×1 convolution mapped each 64-component feature map to the desired class. The network contained a total of 23 convolutional layers. Dense layers were avoided to ensure that the model could process images of any size.

The data pre-processing had included Region of Interest (ROI) cropping, Contrast-Limited Adaptive Histogram Equalization (CLAHE), and pixel resizing. The dataset had then been divided into training, validation, and test sets in a 60:20:20 ratio. To enhance the training process, data augmentation techniques had been applied to the training set, increasing both the number and variability of images.

Following pre-processing, model-specific adjustments had been applied before training. Once trained, the models had been evaluated on the test dataset, and performance metrics such as the Dice score and Jaccard Coefficient had been calculated to assess segmentation accuracy.

The mean Dice scores obtained were 0.958, 0.914, and 93.4 for LV, MYO, and RV, respectively. Additionally, the Hausdorff distances for the proposed method were recorded as 1.69, 2.28, and 1.90 for LV, MYO, and RV, respectively. The p-value for these results was found to be less than 0.05 (=0.0313), indicating the statistical significance of the proposed method.

This automatic, end-to-end trainable computer-based approach required fewer resources and less time while achieving superior results compared to state-of-the-art methods. By improving efficiency, it had the potential to assist medical practitioners in analyzing cardiac diseases more effectively. Given its high performance and statistical significance, this model stands as the current state-of-the-art (SOTA) for cardiac MRI segmentation.

2.7 Comparison of Related Studies

Both traditional image processing/ ML and deep learning approaches have been used for cardiac MRI segmentation of ventricular structures and myocardium, each with its own strengths and weaknesses. This section discusses the strengths and weaknesses of those methods in detail.

2.7.1 Traditional Methods vs. Deep Learning Methods

The evolution of cardiac MRI segmentation techniques can broadly be categorized into two main paradigms: traditional image processing and machine learning (ML)-based methods, and contemporary deep learning (DL)-based approaches. Each paradigm offers a distinct set of advantages and limitations, and their comparative evaluation provides valuable insight into the progression and current state of medical image analysis.

Traditional methods in medical image segmentation typically rely on handcrafted

features derived from domain knowledge, combined with classical image processing techniques or shallow learning algorithms such as support vector machines (SVMs), random forests, or k-nearest neighbors (KNN). These approaches are often appreciated for their **interpretability**, as the decision-making process can be traced back to specific features and heuristics explicitly engineered by researchers or clinicians [29]. Furthermore, these methods are generally **computationally efficient**, requiring significantly less memory and processing power compared to deep neural networks, which makes them suitable for deployment in low-resource settings [29]. Another noteworthy strength lies in their **reduced dependency on large datasets**, since the reliance on explicit feature extraction allows effective model training even with limited annotated data [20].

Despite these advantages, traditional approaches exhibit several limitations. The **manual feature engineering** process is inherently time-consuming and requires deep domain expertise, which can restrict scalability and adaptability [5, 22]. Additionally, the **generalizability** of these methods is often limited. When applied to datasets from different sources, scanners, or imaging protocols, significant reconfiguration of the feature extraction pipeline is often necessary [15]. Moreover, traditional techniques frequently struggle to **capture the complex anatomical variations** present in cardiac structures, particularly in cases involving congenital defects, severe pathology, or motion artifacts [5, 23, 71].

In contrast, deep learning-based methods, particularly convolutional neural networks (CNNs), have revolutionized the field by offering end-to-end learning frameworks capable of automatically learning hierarchical representations from raw image data. These models have demonstrated **state-of-the-art performance** across a wide range of segmentation benchmarks, including cardiac MRI segmentation, often surpassing traditional approaches in both accuracy and robustness [23]. One of the most significant advantages of DL methods is their ability to **automatically extract multi-scale features** without manual intervention [22, 23], reducing the need for handcrafted pipelines and enabling **consistent performance across heterogeneous datasets** [15, 62]. Moreover, their **capacity to model complex spatial relationships** allows them to effectively capture subtle structural variations, making them particularly suitable for segmenting intricate anatomical regions such as the myocardium and cardiac chambers [24].

However, the deployment of deep learning models is not without challenges. These models typically **require large volumes of labeled training data**, which can be scarce in medical imaging due to the need for expert annotations and data privacy constraints [22]. Furthermore, DL models are often criticized for their **lack of interpretability**. The "black-box" nature of neural networks raises concerns in clinical contexts, where explainability is critical for decision support and regulatory approval [72]. Additionally, the **computational demands** of training and deploying DL models can be sub-

stantial, necessitating specialized hardware such as GPUs or TPUs and considerable training time [24].

In summary, while traditional methods offer advantages in interpretability, efficiency, and low data requirements, they fall short in handling the complexity and variability inherent in medical imaging tasks. Deep learning methods, despite their higher resource demands and interpretability concerns, provide a more powerful and scalable solution, particularly when dealing with large and diverse datasets. The ongoing research efforts aim to bridge the gap between these paradigms by incorporating interpretability, efficiency, and data-efficiency into deep models, ultimately enhancing their clinical usability and reliability.

2.7.2 Comparison of Techniques Used for CMRI Segmentation

Various segmentation methods, spanning from no or weak prior approaches to those with strong priors, were employed for cardiac MRI segmentation tasks before the advent of deep learning. Table 2.2, adapted from [20, 29], summarizes the techniques associated with various studies that have used the traditional image processing or ML based methods for CMRI segmentation.

Due to the limitations discussed in Subsection 2.7.1 of the traditional methods, DL-based methods have emerged as powerful tools in Cardiac MRI segmentation tasks. A summary of the findings from previous studies related to CMRI segmentation using DL-based methods is presented in Appendices A.

2.7.3 Comparison of State-of-the-Art Segmentation Methods Evaluated on ACDC Dataset

The performances of the SOTA segmentation methods that were evaluated on the ACDC dataset [18] are summarized in Table 2.3.

Based on the comparison presented in Table 2.4, solution proposed by Chowdary *et al.* [21] demonstrates notable advancements over other existing state-of-the-art (SOTA) methods in cardiac MRI segmentation. While all compared SOTA methods have successfully segmented all three cardiac regions (Left Ventricle, Right Ventricle, and Myocardium) and utilized labeled ACDC data for supervised training, their proposed method distinguishes itself through enhanced generalizability, and superior performance in terms of segmentation accuracy.

2.8 Evaluation Metrics

Accurate segmentation of cardiac structures from CMR images relies on effective loss functions in deep learning models. The choice of a specific loss function depends on the characteristics of the segmentation task and the desired properties of the model

TABLE 2.2: COMPARISON OF TRADITIONAL METHODS USED FOR CMRI SEGMENTATION.

Reference	Category	Technique(s)	Target(s)	Accuracy
[52]	Image-based	Thresholding + edge detection + radial region growth	LV	DC(epi): 0.93 $\pm$ 0.02, DC(endo): 0.89 $\pm$ 0.04
[2]	Image-based	Optimal thresholding + FFT + multiple seeds region growth	LV	DC(epi): 0.94, DC(endo): 0.90
[53]	Image-based + Deformable Method	Region growth with iterative thresholding + active contours	LV	-
[54]	Image-based	Region growth + seeds propagation	LV	-
[3]	Pixel Classification	LV localisation + EM-based classification + active contours	LV	-
[55]	Pixel Classification	GMM (EM) + region restricted dynamic programming	LV	DC(epi):0.94 $\pm$ 0.02, DC(endo): 0.89 $\pm$ 0.03
[56]	Pixel Classification	Geodesic active region + statistical KNN classifier	MYO	DC: 0.79 $\pm$ 0.07
[57]	Pixel Classification + Atlas-based	Multi-atlas + augmented feature + SVM classification	LVM	DC: 0.807
[73]	Deformable Method	Level sets + overlap priors	LV	DC: 0.93 $\pm$ 0.02
[58]	Model-based	ASM + invariant optimal features	LV, RV	-
[59]	Model-based	Hybrid ASM/ AAM	LV, RV	-
[60]	Model-based	ASM + AAM	LV, RV	-

output. Some of the common loss functions used in cardiac MRI segmentation is stated below.

- **Dice Coefficient (DC):** This is a widely used metric to quantify the similarity between the ground truth mask and the predicted mask. It measures the overlap between the two sets, and its values span from 0 to 1, with a perfect agreement indicated by a score of 1 [19, 22]. The mathematical expression for the Dice coefficient in Equation 2.1 is defined as follows, where P and G denote the sets

TABLE 2.3: SEGMENTATION ACCURACY (DICE) OF SOTA METHODS EVALUATED ON ACDC DATASET.

Reference	Method	LV	MYO	RV
Isensee *et al.* [74]	Ensemble model comprising of a 2D and 3D U-Nets	0.950	0.911	0.923
Li *et al.* [75]	ROI detection and segmentation using 2 FCNs	0.944	0.911	0.926
Sun *et al.* [12]	SAUNet - Shape Attentive U-Net	0.938	0.887	0.914
Chowdary *et al.* [21]	Multi-Modal Cardiac Network (MMC-Net)	**0.963**	**0.963**	**0.953**
Yutian *et al.* [5]	Res U-Net as the initial segmentation network and a hierarchical ConvLSTM based recurrent network as the temporal consistency network	0.855	0.746	0.761
da Silva *et al.* [4]	Cascaded approach: UNet for ROI extraction, FCN for initial segmentation, and a U-Net model for refinement	0.938	0.900	0.880
Ren *et al.* [9]	Multi-Task Learning based U-Net (MTLUNET)	0.881	0.807	0.724
Sharan *et al.* [15]	U-Net with VGG encoder and Feature Pyramid Network	0.958	0.914	0.934
Kamal *et al.* [14]	Attention-guided Residual W-Net (ARW-Net)	0.953	0.914	0.923

of pixels contained within the predicted and ground-truth masks, respectively.

$$DC(P,G) = \frac{2 \mid P \cap G \mid}{\mid P \mid + \mid G \mid} \tag{2.1}$$

- **Jaccard Coefficient (JC):** This is commonly known as the Jaccard similarity coefficient and it functions as a metric for gauging the similarity of two sets of pixels. It also measures the extent of overlap between the predicted and ground-truth masks [19, 22]. An elevated Jaccard coefficient value indicates a stronger resemblance between the predicted and ground-truth contours. Calculating the Jaccard coefficients involves a comparison of the sets of pixels encompassed by the predicted (P) and ground-truth (G) label contours, as shown in Equation 2.2.

$$JC(P,G) = \frac{\mid P \cap G \mid}{\mid P \cup G \mid} \tag{2.2}$$

- **Hausdorff Distance (HD):** It measures the degree of mismatch between two sets of points, commonly used to evaluate the spatial accuracy of image segmen-

TABLE 2.4: FEATURE COMPARISON OF THE SOTA METHODS USING ACDC DATASET

Study	**Method**	**All 3 Regions Segmented?**	**Test Set Generaliz-ability**	**Real-time Prediction**
Isensee *et al.* [74]	Ensemble 2D and 3D U-Nets	✓	✗	✗
Li *et al.* [75]	FCNs	✓	✗	✗
Sun *et al.* [12]	SAUNet	✓	✗	✗
Chowdary *et al.* [21]	MMC-Net	✓	✓	✗
Yutian *et al.* [5]	Temporal Consis-tency Network	✓	✗	✗
da Silva *et al.* [4]	Cascaded U-Nets	✓	✗	✗
Ren *et al.* [9]	MTLUNET	✓	✗	✗
Sharan *et al.* [15]	VGG encoder-based U-Net	✓	✗	✗
Kamal *et al.* [14]	ARW-Net	✓	✓	✗

tation results. It is defined as the maximum distance between any point in one set and its nearest point in the other set, considering both directions [19, 22]. A lower HD implies smaller difference between the ground truth and predicted masks. Equation 2.3 shows the HD metric where *d(x, y)* denotes the Euclidean distance between x and y.

$$HD = \max\{\max_{x \in t_i}\{\min_{y \in p_i}[d(x,y)]\}, \max_{y \in p_i}\{\min_{x \in t_i}[d(y,x)]\}\} \tag{2.3}$$

- **Mean Intersection over Union (MIoU):** This method also involves assessing the similarity between the specified ground truth mask and the segmentation result. As shown in Equation 2.4, η_{class} represents the number of classes, n_{ji} signifies the number of pixels classified as class j but originally belonging to class i, and $t_i = \sum_i n_{ii}$ denotes the total number of pixels in class *i*. This method aids in determining the extent of similarity by considering the distribution of pixels across different classes and their origin [10].

$$MIoU = \frac{\sum_i n_{ii}}{\eta_{class} \times (t_i + \sum_j (n_{ji} - n_{ii}))} \tag{2.4}$$

2.9 Limitations and Challenges in Existing Methods

CMRI segmentation of ventricular structures and myocardium is a complex task that involves various challenges and limitations across different segmentation techniques. Both traditional and DL-based segmentation techniques have their own set of challenges in the context of CMRI segmentation.

Many traditional techniques, such as thresholding and region growing, often require manual parameter tuning or user interaction [29, 71]. This subjectivity can lead to variations and may not be suitable for large-scale applications. Moreover, these methods methods may struggle with noisy images or variations in image intensity.

When considering the deformable methods such as active contour and level sets, their performance may heavily depend on the quality of the initial segmentation [18, 29]. These methods may also require user interaction for parameter tuning and guiding the deformable model [19].

The accuracy of atlas-based methods, may depend on the effectiveness of image registration, which may be challenging in the presence of anatomical variations [19, 29]. Also, this method may struggle with abnormal or pathological cases that deviate significantly from the atlas.

The model-based segmentation methods such as SSM, heavily depends on the training data, as the accuracy of SSMs relies on the representativeness and diversity of the training dataset [19]. Unsupervised or supervised techniques, which come under pixel classification methods, also require large amount of training data, and depends on the variations in the input data [19]. Furthermore, the feture engineering can be complex and specific to the image data, and these techniques might struggle with small structures or ambiguous tissue boundaries [29].

Due to above limitations with traditional CMRI segmentation techniques, DL-based techniques have been explored in recent studies; however, those DL-based models require large amounts of labeled data for training, which can be expensive and time-consuming [8, 10, 13]. Moreover, DL models often act as black boxes, making it challenging to understand their decisions [1]. In addition to that, DL-based methods may struggle with generalizing to data outside the training distribution, especially in the presence of domain shifts [10, 14]. Moreover, some of the proposed DL-based architectures are highly customized to specific segmentation task, and might generalize well in other domains [4, 14].

In summary, both traditional and DL-based segmentation techniques face challenges in cardiac MRI segmentation. While traditional methods may struggle with variability and manual parameter tuning, deep learning methods require careful consideration of data annotation challenges, interpretability, and generalization across different datasets. Hybrid approaches that combine the strengths of both techniques may be a promising avenue for addressing these challenges in cardiac MRI segmentation.

CHAPTER 3

METHODOLOGY

U-Net, introduced by Ronneberger et al. in 2015 [6], became a popular architecture for semantic segmentation tasks, particularly in medical image analysis. It has over 98 thousand citations, by the end of 2024. Due to its symmetric and modular design, numerous studies have explored possible modifications that could be done to the original U-Net to improve its performance in segmentation tasks.

Azad *et al.* [24], in their detailed review of medical image segmentation using U-Net variants, have proposed a taxonomy which categorizes the enhancements of the U-Net architecture into six categories such as skip connection enhancements, backbone design enhancements, bottleneck enhancements, transformers, rich representation enhancements and probabilistic design [24].

This study mainly focuses on developing U-Net-based architectures for CMRI segmentation, by combining the ideas and concepts discussed in [24].

3.1 Process Flow

The overall process of cardiac MRI segmentation of ventricular structures and the myocardium consists of data pre-processing, data augmentation, model training and evaluation as depicted in Figure 3.1. The process begins with *data pre-processing*, where the cardiac MRI images are resized to a uniform resolution to ensure consistency across the dataset. The images are then normalized to standardize intensity values, improving the stability of model training. Additionally, de-noising techniques are applied to suppress noise artifacts, enhancing the clarity of the input data for segmentation. Following pre-processing, *data augmentation* techniques are applied to expand the variability within the dataset and improve the model's generalization capabilities. Augmentation operations include rotation, affine transformations, flipping, deformation, and blurring, which mimics real-world variations in imaging conditions and patient anatomy. The pre-processed and augmented data are then fed into the *training phase*, where U-Net variants are trained using a standardized pipeline. The final step involves *model evaluation* using the dice coefficient as the primary metric.

A detailed description of each of these steps will be provided in the upcoming sections.

3.2 Dataset

The Automated Cardiac Diagnosis Challenge (ACDC) dataset, which comprises of cardiac cine MR images, has been compiled using clinical exams conducted at the

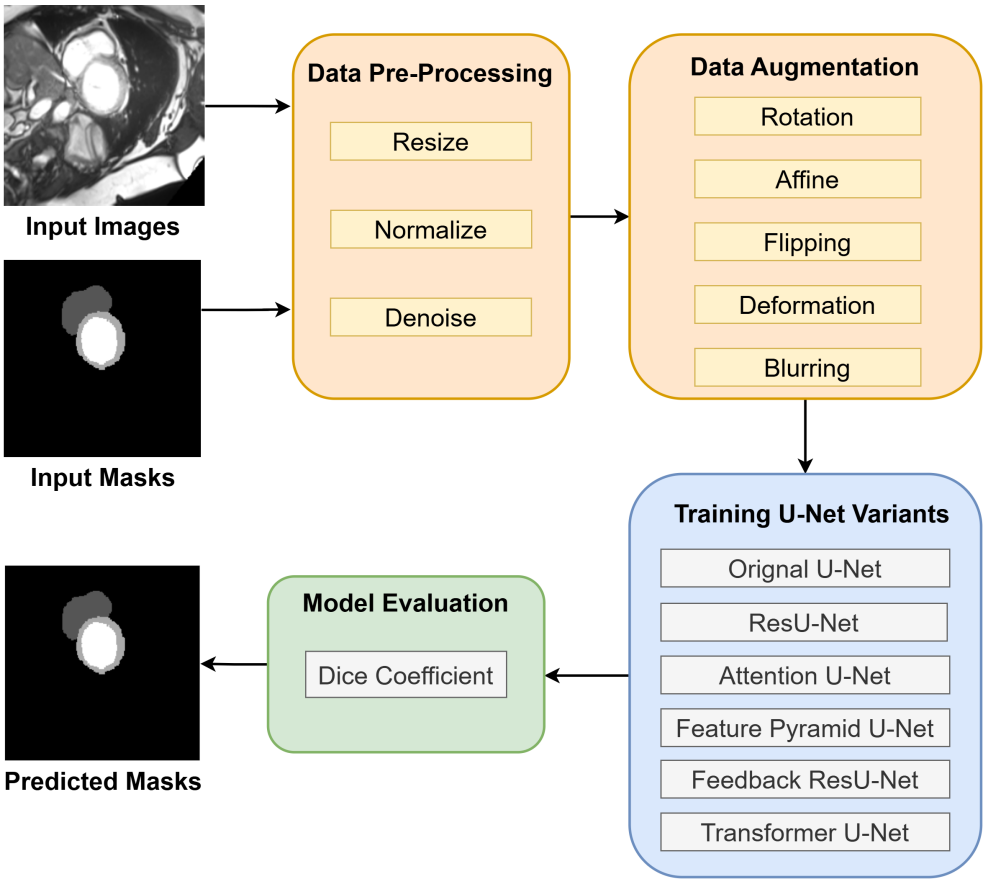

Fig. 3.1: Overall Process of Cardiac MRI Segmentation.

University Hospital of Dijon, France [18]. The Cine Magnetic Resonance (MR) images have been obtained during breath-holding using either retrospective or prospective gating, employing a steady-state free precession (SSFP) sequence in a short-axis orientation [18]. This dataset will be mainly used in this study to develop the U-Net variant for CMRI segmentation of ventricular structures and myocardium.

The ACDC dataset comprises of 150 exams from 150 different patients, and it is divided into 5 balanced subgroups as stated in Table 3.1.

TABLE 3.1: SUMMARY OF THE STUDY POPULATION IN ACDC DATASET

Group	Subgroup	# of Subjects	Label
Pathelogical	Subjects with myocardial infarction	30	MINF
	Subjects with dilated cardiomyopathy	30	DCM
	Subjects with hypertrophic cardiomyopathy	30	HCM
	Subjects with abnormal right ventricle	30	RV
Healthy	Normal subjects	30	NOR

The dataset offers the segmentation ground truths, annotated by clinical experts, for each scan of the LV and RV endocardium and epicardium at both end-diastolic and

end-systolic phases. A sample 2D CMRI image and its ground truth is depicted in Figure 3.2. In this study, we are classifiying each pixel in the input image into one of the 4 classes: Background, LV, RV, or MYO. Every subgroup was distinctly delineated based on physiological parameters, including the LV mass, the local contraction of the LV, left or right diastolic volume or ejection fraction, and the maximum thickness of the myocardium.

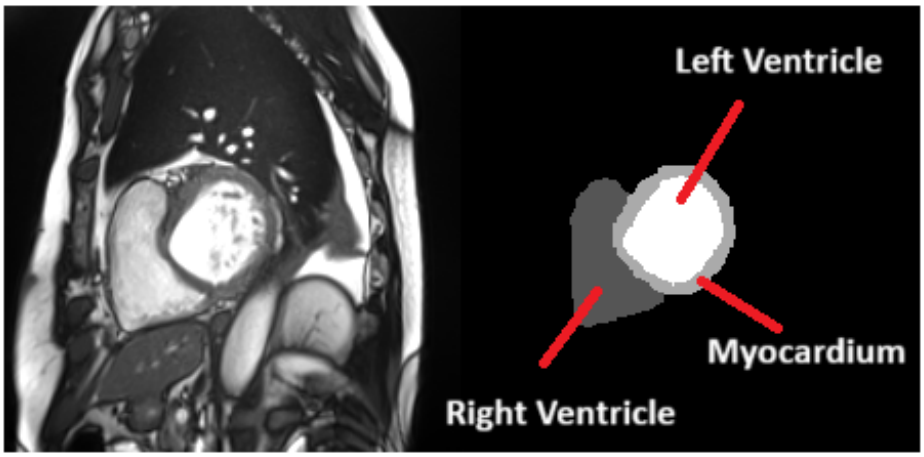

Fig. 3.2: 2D CMR image and its ground truth from the ACDC dataset.

In addition to that, for each scan, the dataset offers additional information about the subject such as age, weight, height and diastolic-systolic phase instants that could be used for further analysis.

The training dataset consists of 100 subjects, with 20 subjects included from each subgroup. Similarly, the test dataset comprises 50 subjects, with 10 subjects from each subgroup.

The structure of the ACDC dataset is shown in Figure 3.3. A description of each of the file present in a patient directory is given below.

- **Info.cfg**: A configuration file comprising the metadata of a patient such as the label (pathology).
- **patientXXX_4d.nii**: The complete MRI sequence in NIfTI format.
- **patientXXX_frameXX.nii**: ED or ES frame in NIfTI format.
- **patientXXX_frameXX_gt.nii**: Annotation at the pixel level for either the ES or ED frame in NIfTI format. The frames are annotated for 3 semantic classes: the left ventricular cavity, the RV, and the Myocardium (MYO) of the LV.

```
acdc_root
  |
  |-----patient001
  |       |-----Info.cfg
  |       |-----patient001_4d.nii
  |       |-----patient001_frame01.nii
  |       |-----patient001_frame01_gt.nii
  |       |-----patient001_frame12.nii
  |       |-----patient001_frame12_gt.nii
  |
  |-----patientXXX
  |       |-----Info.cfg
  |       |-----patientXXX_4d.nii
  |       |-----patientXXX_frameXX.nii
  |       |-----patientXXX_frameXX_gt.nii
  |       |-----patientXXX_frameXX.nii
  |       |-----patientXXX_frameXX_gt.nii
  |
```

Fig. 3.3: The structure of the ACDC dataset.

3.3 Data Pre-Processing

The data preparation and pre-processing steps that are followed are listed below.

- **Data Preparation and Formatting**: The ACDC dataset comprises of ED or ES frames in NIfTI format. 2D image slices are extracted from these NIfTI files, as the study is conducted using 2D images. The 3D images, which are in LPS orientation, were first transformed to RAS orientation. Then, the 2D slices were extracted and stored. The processed repository ACDC directory structure is shown in Figure 3.4. The *images* directory contains the cardiac MRI 2D slices in ".png" format. The *subjectXXX*, *frameXX* and *sliceXX* indicates the subject number, ED or ES frame number, and the 2D slice number in each frame respectively. Moreover, the *masks* directory comprises of binary segmentation masks of LV, RV and MYO in ".png" format. The *labels* directory contains the label (pathology) class of each corresponding subject.

- **Image Resizing**: In order to facilitate the batch processing during model training, images are resized to 224×224 to ensure consistent dimensions across all images and masks. Moreover, the original image anatomical structure was preserved, to avoid distorting anatomical structures.

- **Intensity Normalization**: To enhance the model convergence and stability, pixel values are scaled to a standard range, typically between 0 and 1. This step is crucial as MRI images often have varying intensity ranges. In this study, Z-score normalization was used with 0 mean and a standard deviation of 1.

- **Data Augmentation**: To enhance the model's generalization and robustness capabilities, data augmentation was designed and conducted on the original ACDC 3D images, rather than on 2D slices extracted from them. This approach preserved the spatial relationships within the volumetric data and ensured meaningful variations in the augmented dataset. Augmentation techniques including random rotations, horizontal and vertical flipping, motion blur, affine transformations, and elastic deformations, were applied in a controlled manner. The number of augmented samples was kept within reasonable bounds to avoid generating excessive variations that could lead to incorrect results or overfitting. Furthermore, we conducted experiments to validate the effectiveness of the augmentation strategy by comparing model performance with and without augmented data, ensuring that the augmentation task contributed positively to model learning. Figure 3.5 provides an output of each augmentation technique applied to a sample image.

- **Data Splitting**: The original dataset comprises of training and test datasets with 100 and 50 subjects in each dataset. We have split the total 150 subjects to 80:10:10 ratio and used 120 subjects as training, 15 subjects as validation, and 15 subjects as test data. The 2D image slices extracted from these subjects were used for model training while preserving the 80:10:10 ratio.

- **Noise Reduction**: Image de-noising techniques are considered if the images exhibit significant noise. Median filtering and Gaussian blur were used as noise reduction techniques.

A summary of the image counts for training, validation and test sets before and after applying data augmentation is provided in Table 3.2.

TABLE 3.2: DATASET DISTRIBUTIONS

Dataset	Training	Validation	Test
Original	2160	270	270
After Augmentation	16000	2000	2000

3.4 U-Net Based Variants

The U-Net framework, introduced by Ronneberger *et al.* in 2015 [6], proves to be a robust solution for the segmentation of medical images. In various recent studies, researchers have extensively employed it, adapting its structure to address specific objectives, such as the segmentation of cardiac MRI.

The original U-Net comprises an encoder (contracting) path for extracting features, a decoder (expanding) path for spatial localization and skip connections between the

```
processed_acdc_root
    |
    |-----train
    |        |-----images
    |        |     |-----subject001_frame01_slice02.png
    |        |     |-----subject001_frame12_slice02.png
    |        |     |-----...
    |        |
    |        |-----masks
    |        |     |-----subject001_frame01_slice02_LV.png
    |        |     |-----subject001_frame01_slice02_MYO.png
    |        |     |-----subject001_frame01_slice02_RV.png
    |        |     |-----subject001_frame12_slice02_LV.png
    |        |     |-----subject001_frame12_slice02_MYO.png
    |        |     |-----subject001_frame12_slice02_RV.png
    |        |     |-----...
    |        |-----labels
    |              |-----labels.json
    |
    |-----<dataset>
    |        |-----images
    |        |     |-----subjectXXX_frameXX_sliceXX.png
    |        |     |-----...
    |        |
    |        |-----masks
    |        |     |-----subjectXXX_frameXX_sliceXX_LV.png
    |        |     |-----subjectXXX_frameXX_sliceXX_MYO.png
    |        |     |-----subjectXXX_frameXX_sliceXX_RV.png
    |        |     |-----...
    |        |
    |        |-----labels
    |        |     |-----labels.json
```

Fig. 3.4: The structure of the processed ACDC dataset.

encoder and the decoder levels to bridge the gap between high-level semantic features and low-level spatial details [6].

This study involved making adjustments mentioned below to the original U-Net architecture, and the performance of the modified variants were assessed in the context of segmenting ventricular structures and myocardium in Cardiac MRI.

- **Encoder Enhancements:** Add more encoder levels (i.e. deeper architectures) for effective feature extraction, integrate attention mechanisms to focus on relevant features for specific structures, or introduce residual connections with encoder blocks for improved information flow and gradient propagation [24].
- **Skip Connection Enhancements:** Use attention mechanism to process the feature maps within the skip connections.
- **Bottleneck Enhancements:** Use attention mechanism in the bottleneck of the U-Net.
- **Transformers:** Develop a customized and hybrid design Transformer model which includes patch embedding and multi-head self-attention mechanism [76, 77].

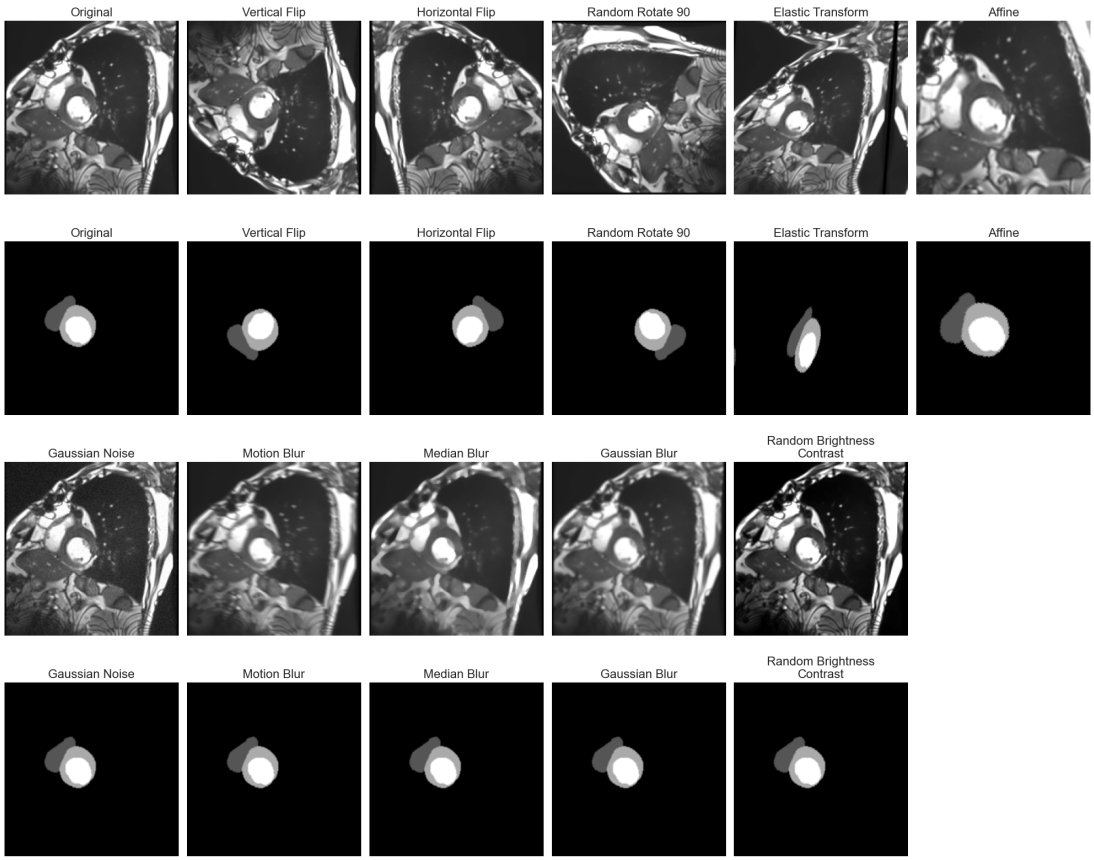

Fig. 3.5: Image and Mask Augmentation.

- **Rich Representation Enhancements:** Use multi-scale (pyramid method) in the encoder path, which resizes the input image and into a set of decreasing spatial resolution images [24] and fuse those as inputs at different encoder levels.
- **Feedback Mechanism:** Incorporate feedback of the decoder output to the encoder input to iteratively refine the segmentation output [7].

In the following sub-sections, the U-Net variants that were developed are discussed in detail.

3.4.1 Original U-Net (O-UN)

The Original U-Net (O-UN) architecture proposed by Ronneberger *et al.* [6] is the baseline model used in this study to compare the performances of U-Net variants. It follows a symmetric encoder-decoder structure with skip connections, which preserve spatial data lost during downsampling. The encoder extracts features via a series of convolutional and max-pooling layers (refer Equation 3.1 and Equation 3.2), while the decoder performs upsampling to reconstruct the segmented output. In Equation 3.1 $*$ denotes convolution, x is the input feature map, W is the kernel, b is bias, and f is the

activation function (e.g., ReLU). Skip connections directly pass feature maps from the encoder to the corresponding decoder layers, enabling precise localization and efficient feature reuse as provided in Equation 3.3.

$$y = f(W * x + b) \tag{3.1}$$

$$y_{i,j} = \max_{(m,n)\in\mathcal{N}(i,j)} x_{m,n} \tag{3.2}$$

$$z = x_{\text{encoder}} \oplus x_{\text{decoder}} \tag{3.3}$$

The U-Net architecture is composed of two main parts as shown in Figure 3.6.

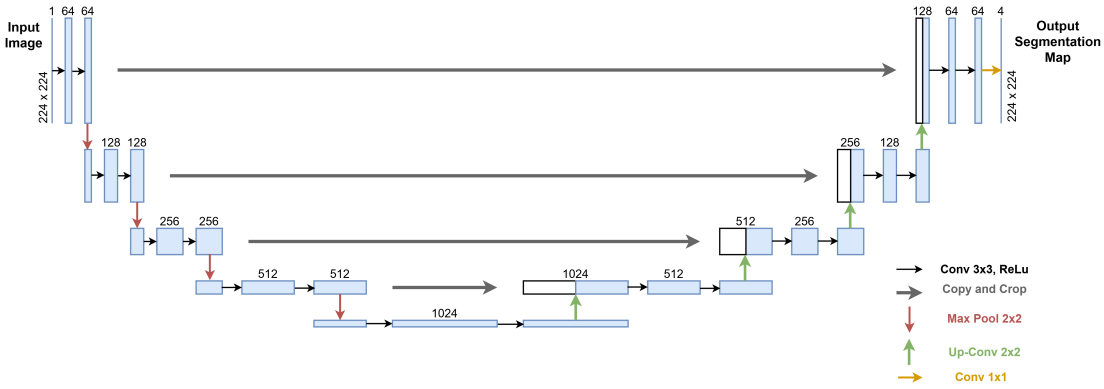

Fig. 3.6: The Original U-Net Architecture.

The **contracting path (encoder)** follows the structure of a typical convolutional neural network. It consists of repeated blocks of two 3×3 convolutional layers (each followed by a ReLU activation function) and a 2×2 max-pooling layer for downsampling. Each downsampling step doubles the number of feature channels, allowing the network to learn more complex features as the spatial dimensions reduce. The encoder captures high-level features and context information about the input image.

On the other hand, the **expansive path (decoder)** performs upsampling and feature reconstruction. Each upsampling step is achieved using a 2×2 transposed convolution (or deconvolution) layer, which halves the number of feature channels while increasing spatial resolution. The upsampled feature map is concatenated with the corresponding feature map from the contracting path using skip connections. These skip connections provide fine-grained spatial details from earlier layers, helping to recover lost spatial information due to downsampling. After concatenation, the features are passed through two 3×3 convolutional layers with ReLU activations.

The **final output layer** is a 1×1 convolution that maps the features to the desired number of classes for segmentation, which is 4 in our study. This produces a pixel-wise classification of the input image.

3.4.2 Residual U-Net (Res-UN)

The Residual U-Net (Res-UN) , also known as ResU-Net, adapted from [16], combines the strengths of U-Net for segmentation tasks with the benefits of residual learning, which helps in training deep networks by addressing vanishing gradient issues. The 5 key components of the architecture, as depicted in Figure 3.7 are: residual convolutional block, encoder, bottleneck, decoder and the output layer.

Each **residual convolutional block** in this architecture uses two convolutional layers with kernel size 3 and padding 1 for spatial consistency. The batch normalization normalizes the input features across the batch dimension for each channel, while the ReLu activation function is used for introducing non-linearity. The residual connection F(x), as in Equation 3.4, bypasses the main path using either an identity mapping or a 1×1 convolution (if input and output channel dimensions differ) [16]. This connection allows the network to learn modifications to the input rather than entirely new representations, improving gradient flow during training [16]. x and y in Equation 3.4 refers to input and output tensors while σ refers to ReLu activation. Conv_1 and Conv_2 refer to 3×3 convolution operations, while BN_1 and BN_2 refer to Batch Normalization.

$$\mathbf{y} = \sigma\left(\text{BN}_2\left(\text{Conv}_2\left(\sigma\left(\text{BN}_1\left(\text{Conv}_1(\mathbf{x})\right)\right)\right)\right)\right) + \mathcal{F}(\mathbf{x}) \tag{3.4}$$

While the original ResU-Net architecture uses 3 consecutive residual blocks [16], a fourth residual block was added to the encoder to deepen the architecture and as a novelty as shown in Figure 3.7. Therefore, the **encoder** composes of four consecutive residual blocks, progressively increasing the number of channels from 64 to 512. Each block is followed by a max-pooling layer for spatial down-sampling, reducing the input size and extracting hierarchical features.

As the **bottleneck** layer, another residual block is used with 1024 channels. It acts as a feature aggregation layer, capturing the most abstract and high-level representations of the input [16].

Each **decoder** level consists of an upsampling operation to restore the spatial resolution, skip connections from the corresponding encoder level, implemented using concatenation, and a residual block for further feature refinement as in Figure 3.7. This design ensures that the high-resolution features from the encoder are reused, enhancing spatial detail preservation [16].

At last, the **output layer** comprises of a convolutional layer that reduces the number of channels to the required number of output classes, which is 4 in our study. This produces the segmentation map for the input image.

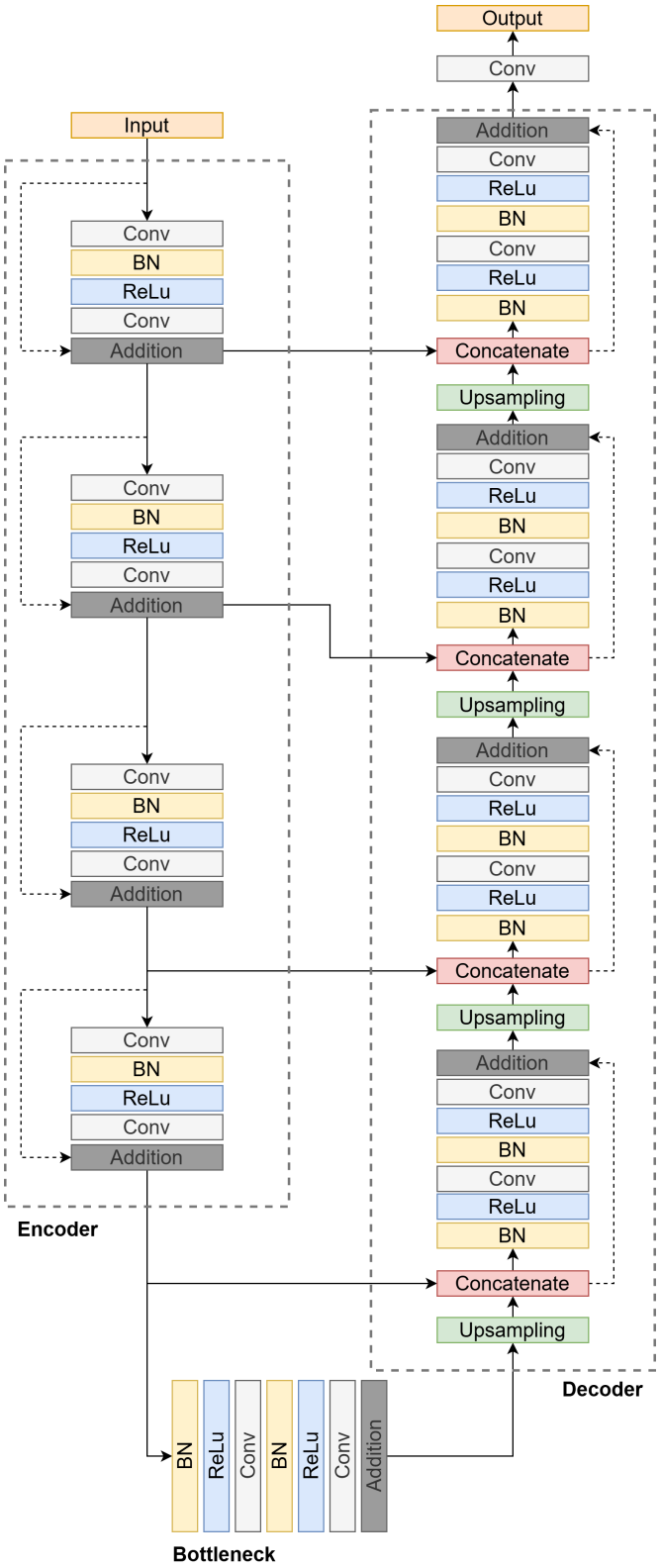

Fig. 3.7: The Residual U-Net Architecture. Adapted from [16].

3.4.3 Attention U-Net (Atn-UN)

The Attention U-Net (Atn-UN) developed in this study is an advanced version of the conventional U-Net, which introduces attention mechanisms to improve the segmentation performance by focusing on the most relevant spatial features [11]. The **attention block** used in this architecture refines the features passed between the encoder and decoder paths. This block calculates attention coefficients to suppress irrelevant regions and highlight salient features. As shown in Figure 3.8, the attention block comprises of below components.

- **Gating and Linear Transformations:** Separate linear transformations are applied to the input feature maps g and the skip connection feature maps x^l, reducing their dimensions for computational efficiency (refer Equation 3.5). W_x and W_g are the learnable weights ($1 \times 1 \times 1$ convolutions)
- **Additive Combination:** The outputs from these transformations are added, followed by a ReLU activation (refer Equation 3.6). σ_1 refers to ReLu activation while f denotes the intermediate feature map.
- **Sigmoid Attention Map:** The combined output is passed through a $Sigmoid$ function to generate the attention weights (refer Equation 3.7). ψ is another $1 \times 1 \times 1$ convolution and σ_2 is the Sigmoid activation. α denotes the attention coefficient map.
- **Feature Recalibration:** The skip connection features are multiplied by the attention weights to emphasize relevant areas (refer Equation 3.8). $\odot$ refers to element-wise multiplication (resampling or gating).

$$\theta_x = W_x * x^l, \quad \phi_g = W_g * g \tag{3.5}$$

$$f = \sigma_1(\theta_x + \phi_g) \tag{3.6}$$

$$\alpha = \sigma_2\left(\psi * f\right) \tag{3.7}$$

$$\hat{x}^l = \alpha \odot x^l \tag{3.8}$$

In addition to that, each decoder and encoder block uses a double convolution block with batch normalization and ReLU activations, which extracts hierarchical features efficiently. Figure 3.9 shows the proposed novel Attention U-Net architecture with an additional encoder layer.

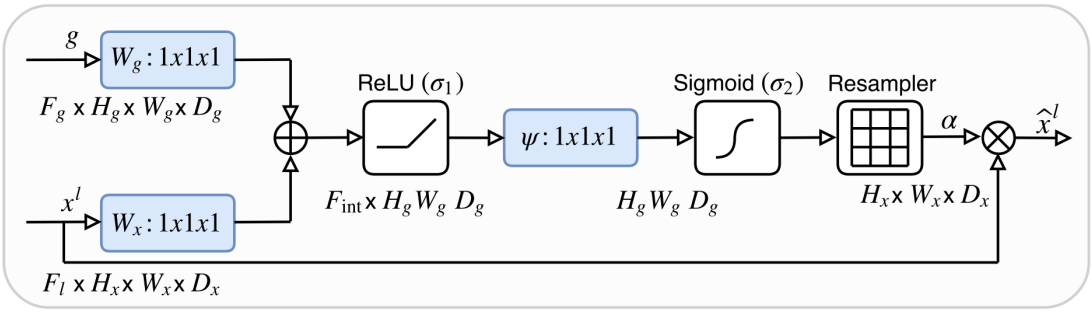

Fig. 3.8: The Attention Block. Adapted from [11].

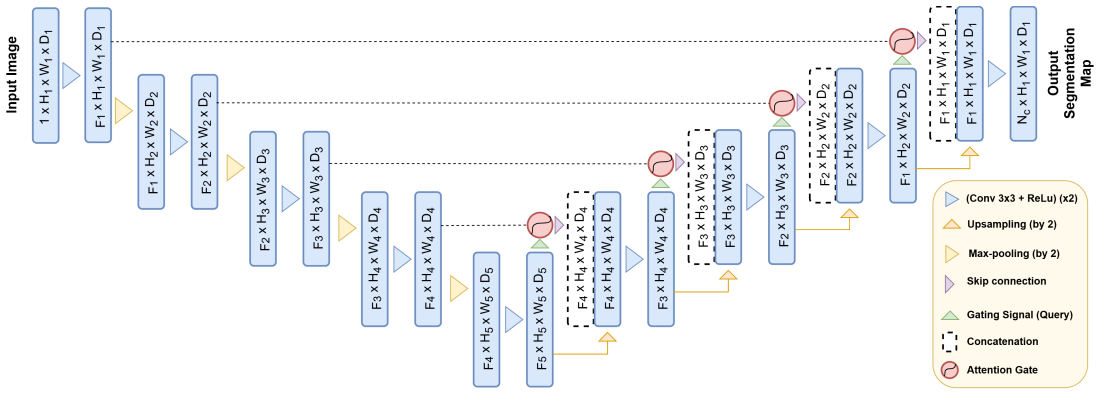

Fig. 3.9: The Attetion U-Net Architecture

In the input and encoder path, the input image first passes through multiple encoder blocks, where each block consists of two convolutional layers followed by a pooling operation. Feature maps are progressively downsampled, capturing high-level semantic information.

At the bottleneck, which is the deepest layer in the U-Net structure, it aggregates global features. Attention is applied at the bottleneck to focus on the most significant global features before passing them to the decoder.

The decoder path consists of upsampling layers followed by concatenation with corresponding encoder features. Before concatenation, attention blocks refine the encoder features to include only the most relevant spatial information. Moreover, it uses double convolution blocks which refines the upsampled features.

The skip connections used in the architecture are not directly concatenated. Instead, they are first processed through attention blocks to remove redundant or less important features.

3.4.4 Feature Pyramid U-Net (FP-UN)

The Feature Pyramid U-Net (FP-UN) is a modified version of the U-Net architecture designed to enhance feature extraction and multi-scale information integration through its novel Feature Pyramid Block. The architecture diagram of the proposed FP-UN is

depicted in Figure 3.10.

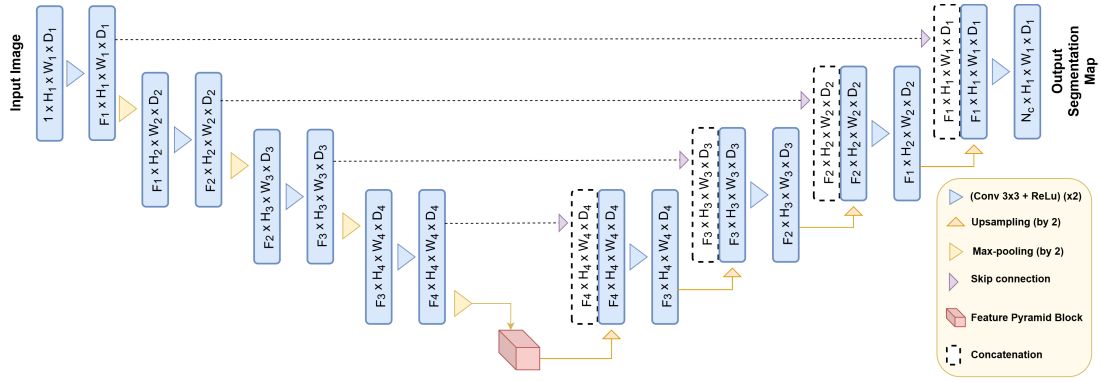

Fig. 3.10: The Feature Pyramid U-Net Architecture

In the **encoder path**, the encoder blocks reduce the spatial dimensions of the input while increasing the number of feature channels. This is achieved through sequential convolutional blocks and pooling operations. Each convolutional block consists of two 2D convolutional layers with a kernel size of 3×3 followed by Batch Normalization and ReLu activation. These blocks learn hierarchical features at increasing levels of abstraction. After each convolutional block, a max-pooling operation down-samples the feature map by a factor of 2, reducing spatial dimensions while retaining key features.

The **Feature Pyramid Block (FPB)** is the unique component of this architecture, enabling multi-scale feature learning to capture context at various spatial resolutions. It comprises of below components as shown in Figure 3.11.

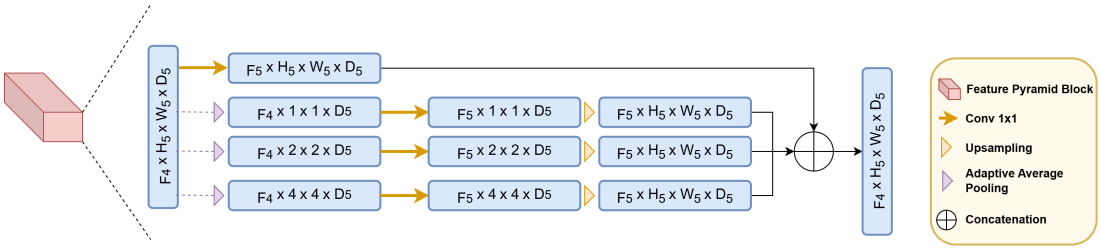

Fig. 3.11: The Feature Pyramid Block

- **Input Transformation:** A 1×1 convolution reduces the input feature map into a lower-dimensional space for computational efficiency (refer Equation 3.9). W_1 is the learnable kernel.

- **Multi-Scale Feature Extraction:** The FPB uses adaptive average pooling to create smaller feature maps at three scales: 1×1 pooling captures global information by compressing the entire feature map into a single value per channel, 2×2 pooling extracts medium-scale context, and 4×4 pooling focuses on finer

details while maintaining spatial coherence. Each pooled feature map undergoes a 1×1 convolution to learn new representations (refer Equations 3.10 - 3.12).

- **Upsampling:** The pooled feature maps are upsampled back to the original spatial dimensions using bilinear interpolation. This ensures that the multi-scale features align spatially with the original feature map (refer Equations 3.13 - 3.15).
- **Concatenation:** The upsampled feature maps from all scales, along with the original transformed feature map, are concatenated along the channel dimension to create a rich, multi-scale representation (refer Equation 3.16).

$$X' = W_1 * X \tag{3.9}$$

$$P_1 = \text{Conv}_{1\times1}(\text{AvgPool}_{1\times1}(X')) \tag{3.10}$$
$$P_2 = \text{Conv}_{1\times1}(\text{AvgPool}_{2\times2}(X')) \tag{3.11}$$
$$P_3 = \text{Conv}_{1\times1}(\text{AvgPool}_{4\times4}(X')) \tag{3.12}$$

$$U_1 = \text{Upsample}(P_1, \text{size} = H \times W) \tag{3.13}$$
$$U_2 = \text{Upsample}(P_2, \text{size} = H \times W) \tag{3.14}$$
$$U_3 = \text{Upsample}(P_3, \text{size} = H \times W) \tag{3.15}$$

$$F_{\text{fpb}} = \text{Concat}(X', U_1, U_2, U_3) \tag{3.16}$$

In the **decoder path**, the decoder reconstructs the image from the compressed multi-scale feature representation produced by the FPB. It progressively restores spatial resolution using transposed convolutions and integrates fine details from the encoder via skip connections. The upsampling and decoding involves transposed convolutions where each decoding step doubles the spatial dimensions. These upsampled feature maps are then concatenated with corresponding encoder feature maps. This preserves spatial details lost during downsampling. After concatenation, another convolutional block (similar to the encoder's convolution block) processes the combined feature maps.

Finally, the **output layer** uses 1×1 convolution layer reduces the channel count of the final feature map to match the number of target classes, which is 4 in our study. This produces pixel-wise class probabilities for segmentation.

3.4.5 Feedback Residual U-Net (Feed-Res-UN)

The Feedback Residual U-Net (Feed-Res-UN) proposed in this study is an enhanced version of the standard U-Net, which incorporates feedback mechanisms into its design. This architecture, as shown in Figure 3.12, is particularly useful in segmentation tasks, such as medical image segmentation, where preserving fine details and incorporating previous output into the learning process can improve the overall model performance [7].

Similar to the Residual U-Net, Feed-Res-UN comprises of **Residual Blocks**, which are the core building blocks of the network. They incorporate residual connections to help the model learn residual mappings. This improves the gradient flow and aids in training deep networks by mitigating vanishing gradient problems. Each residual block consists of two convolutional layers with batch normalization. A skip connection is included to match the input and output dimensions when the number of channels in the residual connection differs from the current block.

The **encoder** in the Feed-Res-UN consists of a series of residual blocks that progressively downsample the input image. The encoder reduces the spatial dimensions while increasing the depth of the feature maps. The first four residual blocks are used for downsampling, and after each block, a max pooling operation is applied to reduce the spatial dimensions.

The **bottleneck** consists of a Residual Block with a large number of output channels (1024 in our case), which enables the network to capture complex features from the downsampled representation before upsampling begins.

The **decoder** consists of a series of upsampling operations (using transposed convolutions) to progressively reconstruct the original spatial dimensions of the input image. Each upsampling operation is followed by a concatenation with the corresponding encoder feature map, which is a key part of the standard U-Net architecture. This skip connection ensures that the decoder has access to high-resolution features from the encoder, helping the network recover fine-grained details. After each concatenation, a residual block processes the combined feature maps.

A unique feature of the Feedback Residual U-Net is its **feedback mechanism**, which introduces the output of the first round of processing back into the network as additional input. After the first forward pass, the output is passed through a 1×1 convolution to adjust its channel dimensions to match the input. The adjusted output is then added to the original input image, creating a feedback loop. The modified input (now including the feedback from the first pass) is processed through the network again in a second round. The whole feedback process is defined in Equations 3.17 - 3.20. The X denotes the original input and $F(.)$ denotes the U-Net processing function as a whole. $Y^{(1)}$ denotes the output after the first forward pass, where $\phi(.)$ is 1×1 convolution to match the input dimensions. This feedback helps refine the model's

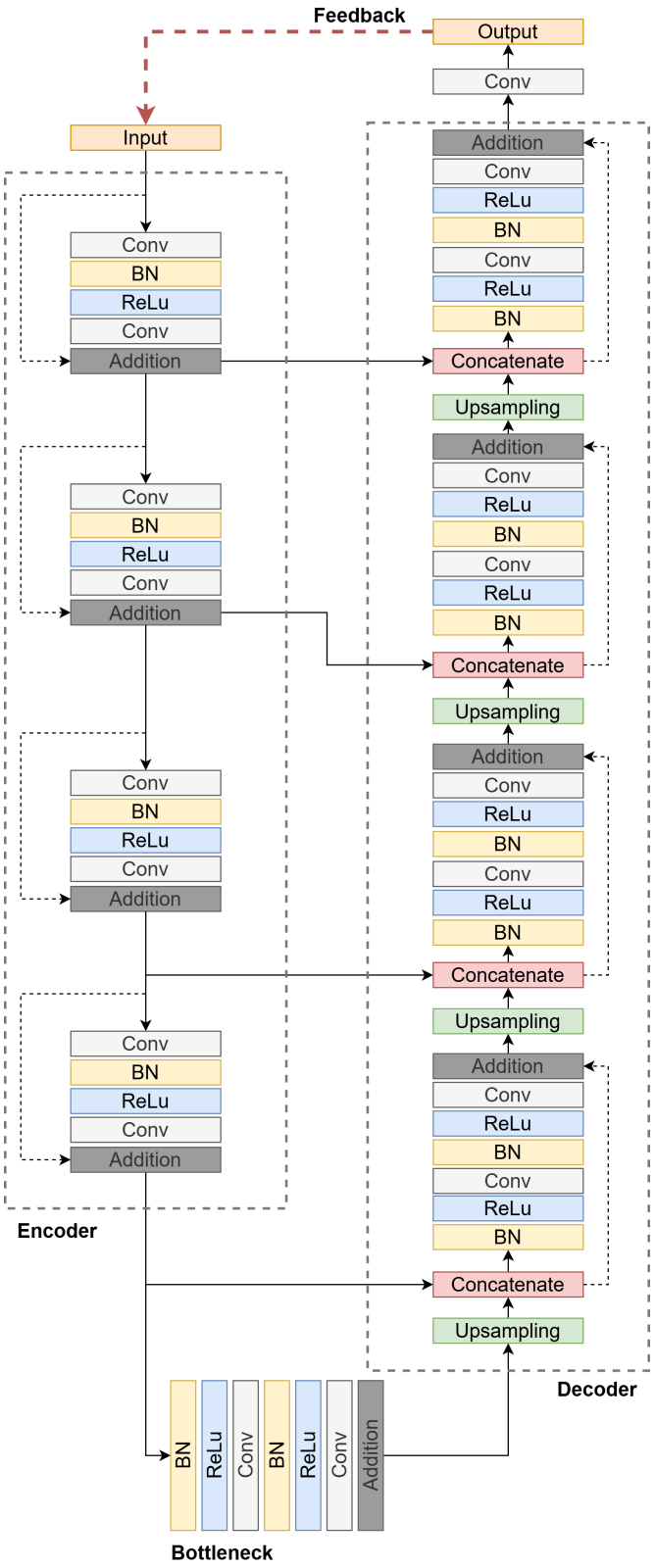

Fig. 3.12: The Feedback Res U-Net Architecture

predictions by incorporating previous output into the current learning step.

$$Y^{(1)} = \mathcal{F}(X) \tag{3.17}$$
$$\hat{X} = X + \phi(Y^{(1)}) \tag{3.18}$$
$$Y^{(2)} = \mathcal{F}(\hat{X}) \tag{3.19}$$
$$\hat{Y} = \psi(Y^{(2)}) \tag{3.20}$$

After the second round of processing, a final 1×1 convolution is applied to produce the output segmentation map. This segmentation map corresponds to the predicted labels for each pixel, which represents the classes such as background, LV, RV and MYO.

3.4.6 Transformer-Based U-Net (Trans-UN)

Combining the ideas present in TransUNet [77] and ViT [76] studies, a novel transformer-based architecture was developed and trained in this study. The architecture diagram of the proposed model is shown in Figure 3.13.

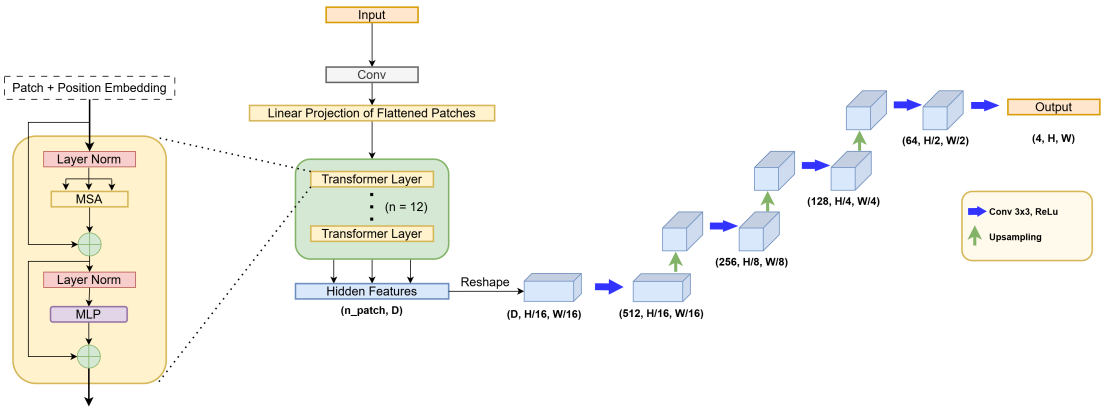

Fig. 3.13: The Transformer-Based U-Net Architecture

The architecture comprises of 3 main parts: patch embedding module, transformer encoder and the U-Net decoder. In the **Patch Embedding** module, the input image is divided into fixed-size patches and each patch is embedded into a high-dimensional vector. This is done using the initial convolution layer, and each generated patch is then projected into a feature space of specified embedding dimension. A positional embedding is added to the patch embeddings to retain spatial information lost during patch tokenization.

The **Transformer Encoder** extracts global context and dependencies across the entire image, as the next step. The transformer encoder comprises of below key components.

- **Multi-Head Attention:** Captures interactions between all patch embeddings.
- **Feed-Forward Network (FFN):** Enhances feature transformation via a two-layer Multi-Layer Perceptron (MLP) .
- **Layer Normalization:** Ensures stability and faster convergence during training.
- **Dropout:** Adds regularization to prevent overfitting.
- **Stacked Layers:** Multiple encoder blocks are stacked to increase the representation power.

The **U-Net Decoder** reconstructs the spatial resolution and refines the segmentation details. It comprises of convolutional blocks that include two convolutional layers followed by ReLU activation in each block and upsaling layers that uses transpose convolutions that progressively upscale the feature maps.

3.5 Loss Function

A custom hybrid loss function was employed in this study, as another contribution of this study, to effectively optimize the segmentation of ventricular structures and myocardium in cardiac MRI images. The custom loss combines two complementary components: *Cross-Entropy Loss* and *Dice Loss*.

The **Cross-Entropy Loss** (L_{CE}), provided in Equation 3.21, is a widely used loss function for multi-class classification tasks. It measures the difference between the predicted probability distribution (softmax outputs) and the ground-truth labels.

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{3.21}$$

The **Dice Loss** (L_{Dice}), is specifically designed for segmentation tasks to address class imbalance and focus on overlapping regions between the predicted and ground-truth masks. The dice loss calculates a soft dice coefficient between the predicted softmax probabilities and the one-hot encoded ground truth labels. This metric is particularly effective for capturing the similarity between the segmented output and the target, especially in cases where certain classes occupy a small fraction of the image. For multi-class segmentation, the dice loss is computed as given in Equation 3.22.

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N}\hat{y}_{i,c}y_{i,c}}{\sum_{c=1}^{C}\sum_{i=1}^{N}(\hat{y}_{i,c}^2 + y_{i,c}^2)} \tag{3.22}$$

In Equation 3.21, N refers to total number of pixels, and C refers to the number of classes, which is four in our study including the background. $\hat{y}_{i,c}$ in Equation 3.21 and

Equation 3.22 refers to the predicted probability for pixel i and class c, and $y_{i,c}$ refers to the ground truth (one-hot encoded) for pixel i and class c.

To leverage the strengths of both loss functions, a weighted combination of Cross-Entropy Loss and Dice Loss was employed, with the weights 0.6 and 0.4 determined experimentally to achieve optimal performance, as provided in Equation 3.23.

$$\mathcal{L}_{Custom} = 0.6 \cdot \mathcal{L}_{CE} + 0.4 \cdot \mathcal{L}_{Dice} \tag{3.23}$$

3.6 Experimental Setup

The following tools and libraries were employed to facilitate the preparation and pre-processing of cardiac MRI data, ensuring compatibility with the implemented machine learning models and robust handling of medical imaging data:

- **TorchIO:** TorchIO was employed in this study to handle 3D cardiac MRI data efficiently, enabling advanced pre-processing and augmentation. Its capabilities include spatial transformations like resampling and cropping, intensity standardization through normalization and histogram matching, and augmentations such as random flipping, rotation, and elastic deformations. TorchIO's focus on volumetric data ensured the effective preparation of high-dimensional MRI datasets while enhancing the generalizability of the models through robust augmentation techniques.

- **NiBabel:** NiBabel is a dedicated library for handling neuroimaging file formats like NIfTI, widely used for medical imaging. It was used in this study for reading cardiac MRI 3D images, and extrcting 2D images slices from them.

- **Albumentations:** Albumentations is a fast and flexible library for image augmentation, widely used in computer vision tasks to improve model generalization. In this study, Albumentations was employed to augment 2D slices extracted from 3D cardiac MRI data, enriching the training dataset with a diverse range of variations. Its powerful augmentation techniques, including geometric transformations (flipping, rotation, scaling), intensity modifications (brightness and contrast adjustments), and noise addition, were instrumental in simulating real-world variability.

The model training process was configured with a robust setup to ensure optimal performance and generalization. The model was trained using **PyTorch** 2.2.1 framework on an NVIDIA GeForce RTX 2080 Super GPU, with a batch size of 16 to balance memory efficiency and gradient stability. An initial learning rate of 5e-4 was employed, with dynamic adjustments managed by the ReduceLROnPlateau scheduler,

which reduces the learning rate when the validation loss plateaus, facilitating fine-tuned optimization during later stages of training.

The optimization process utilized the AdamW optimizer, an improved variant of Adam that includes decoupled weight decay regularization, which mitigates overfitting and ensures better generalization by separately penalizing large weights. Training was conducted for a maximum of 50 epochs, with early stopping implemented to halt training when the validation performance ceased to improve, reducing the risk of overfitting. This setup ensured a computationally efficient training process while preserving model accuracy.

3.7 Web Application Development

To enhance the accessibility and usability of the cardiac MRI segmentation model, a web-based application was developed and deployed on the **Hugging Face Spaces** platform. The primary objective of this web app is to bridge the gap between complex deep learning models and clinical or educational use by providing an intuitive interface for real-time image analysis. Such an application has the potential to significantly benefit clinicians, radiologists, and medical researchers who may not have the technical expertise or computational resources to run segmentation algorithms locally. The application is publicly hosted and accessible at the following Hugging Face URL: cmri-segmentation-web-app.

Built using the **Gradio** framework, this web application integrates a trained segmentation model, specifically the best performing Feature Pyramid UNet in this study, to automatically identify and highlight critical anatomical structures in cardiac MRI slices, including the Left Ventricle, Right Ventricle, and Myocardium. Users can simply upload a cardiac MRI image, trigger the segmentation with a single click, and instantly receive an overlaid output highlighting the segmented regions.

Figure 3.14 and Figure 3.15 illustrate the user interface of the developed web application for cardiac MRI segmentation. Figure 3.14 presents the initial state of the interface, where the user is prompted to upload a cardiac MRI image via drag-and-drop or file selection. The layout is clean and logically divided into functional sections: the left panel is reserved for the input image, while the right panel displays the segmented output. The interface also features “Submit” and “Download Output” buttons for processing and retrieving results, along with a “Region Description” textbox that dynamically updates based on user interaction. Below these, a set of example images is made available for instant testing of the application without the need for external uploads. Moreover, a “Clear” button is used to reset inputs and outputs of the web application.

Figure 3.15 demonstrates the application in action following the segmentation of a sample cardiac MRI. The output image visually highlights the segmented heart struc-

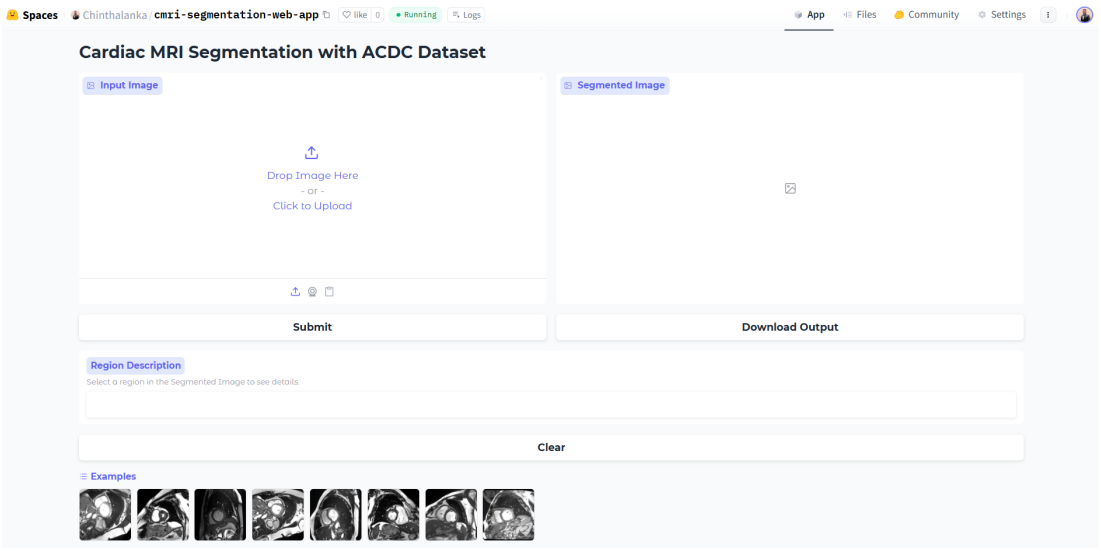

Fig. 3.14: The Web Application

tures: Left Ventricle (red), Right Ventricle (blue), and Myocardium (green), superimposed on the original grayscale scan. This visual distinction allows for immediate anatomical interpretation. Additionally, upon selecting a segmented region, the "Region Description" field populates with a concise explanation of the selected anatomical structure, thereby enhancing the educational and clinical value of the tool. In the example shown, the Left Ventricle has been selected, and its functional description is displayed. This contextual annotation offers clinicians a quick reference, which can be particularly helpful for trainees or interdisciplinary teams.

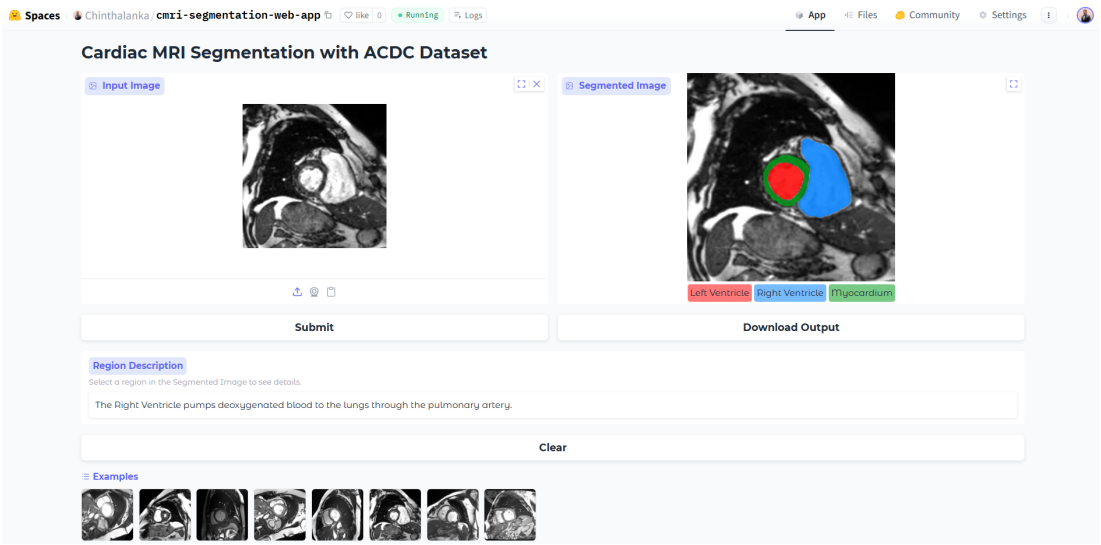

Fig. 3.15: Sample Image Segmented Using the Web Application

This web-based deployment underscores the growing need for AI-assisted tools that are not only powerful, but also readily deployable in real-world medical envi-

ronments. By offering a plug-and-play solution that can be used on any device with a browser, this application represents a significant step towards integrating machine learning-driven diagnostic support tools into routine clinical workflows. Such tools have the potential to accelerate diagnosis, reduce human error, and support a more standardized interpretation of cardiac imaging data.

CHAPTER 4

RESULTS

This chapter presents the results obtained in the Cardiac MRI segmentation experiments conducted using different U-Net variants.

4.1 Evaluation Metrics

The Dice Coefficient (DC) in Equation 4.1, was used as the primary evaluation metric to assess the performance of different U-Net variants. This metric is commonly used in medical image analysis, including tasks like cardiac MRI segmentation, because it provides a quantitative measure of how closely the predicted segmentation aligns with the ground truth. It measures the overlap between the two sets, and its values span from 0 to 1, with a perfect agreement indicated by a score of 1. P and G in Equation 4.1 denote the sets of pixels contained within the predicted and ground-truth masks, respectively.

$$DC(P, G) = \frac{2 \mid P \cap G \mid}{\mid P \mid + \mid G \mid} \tag{4.1}$$

The Jaccard Coefficient (JC) in Equation 4.2 was used as the secondary metric to evaluate the performance of the U-Net variants. This metric, also known as the Intersection over Union (IoU), is also widely utilized in medical image analysis, particularly in segmentation tasks such as cardiac MRI segmentation. It offers a quantitative measure of the similarity between the predicted segmentation and the ground truth by calculating the ratio of the intersection to the union of the predicted and actual masks. The Jaccard Index ranges from 0 to 1, where a value of 1 indicates perfect agreement between the predicted and ground-truth segmentations. In Equation 4.2, P and G represent the sets of pixels in the predicted and ground-truth masks, respectively.

$$JC(P, G) = \frac{|P \cap G|}{|P \cup G|} \tag{4.2}$$

4.2 ACDC Test Set Performance

The Dice Scores obtained by each U-Net model for the LV, RV and MYO regions in the test set are summarized in Table 4.1. All six models have shown better results in segmenting ventricular structures and myocardium in CMRI images.

The results in Table 4.1 indicate following key insights:

- The **Original U-Net (O-UN)** showed robust performance across all categories. It achieved a Dice score of 0.9411 for the Left Ventricle (LV), 0.8911 for the

TABLE 4.1: EVALUATION OF U-NET VARIANTS USING TEST SET DICE SCORES

Model	LV	RV	MYO
Original U-Net (O-UN)	0.9411	**0.8911**	0.8511
Residual U-Net (Res-UN)	0.9219	0.8449	0.8204
Attention U-Net (Atn-UN)	0.9374	0.8810	0.8403
Feature Pyramid U-Net (FP-UN)	**0.9445**	0.8883	**0.8513**
Feedback Residual U-Net (Feed-Res-UN)	0.9389	0.8896	0.8455
Transformer-Based U-Net (Trans-UN)	0.8679	0.8013	0.7191

Right Ventricle (RV), and 0.8511 for the Myocardium (MYO). Notably, it recorded the highest Dice score for RV segmentation among all tested models, highlighting its robustness in capturing the complex structure of the right ventricle. Despite the emergence of advanced architectures, the Original U-Net remained highly competitive, providing a strong baseline for comparison with other enhanced variants.

- The **Residual U-Net (Res-UN)**, which incorporated residual connections to facilitate better gradient flow, underperformed compared to the Original U-Net across all regions. It achieved Dice scores of 0.9219 for LV, 0.8449 for RV, and 0.8204 for MYO. These results suggest that while residual learning can be beneficial in deeper networks, in this case, it may have introduced unnecessary complexity without significant gains in segmentation accuracy. Consequently, the Residual U-Net did not offer substantial improvement over the baseline for cardiac MRI segmentation.

- The **Attention U-Net (Atn-UN)** demonstrated moderate improvements over the Residual U-Net, achieving Dice scores of 0.9374 for LV, 0.8810 for RV, and 0.8403 for MYO. The integration of attention gates helped the model focus on relevant features during segmentation, leading to better performance, particularly for RV and MYO, compared to the residual variant. However, the Attention U-Net still fell slightly short of the Original U-Net's performance, indicating that while attention mechanisms were helpful, they were not sufficient on their own to outperform the baseline model in this task.

- The **Feature Pyramid U-Net (FP-UN)** emerged as the best-performing variant overall, delivering Dice scores of 0.9445 for LV, 0.8883 for RV, and 0.8513 for MYO. It achieved the highest scores for both LV and MYO segmentation tasks and demonstrated strong competitive performance for RV as well. By integrating multi-scale feature extraction through a feature pyramid structure, the model effectively captured the varying anatomical sizes and complexities present in

cardiac structures. This enhanced ability to manage spatial hierarchies made FP-UN particularly suitable for cardiac MRI segmentation.

- The **Feedback Residual U-Net (Feed-Res-UN)** improved upon the standard Residual U-Net by incorporating feedback mechanisms from the decoder to the encoder, resulting in Dice scores of 0.9389 for LV, 0.8896 for RV, and 0.8455 for MYO. The model demonstrated competitive performance, especially in RV segmentation, closely trailing the Original U-Net and Feature Pyramid U-Net. The feedback connections appeared to help refine the feature representations, leading to better segmentation results than the plain residual design, although it still did not surpass the best-performing models.

- The **Transformer-Based U-Net (Trans-UN)** exhibited the weakest performance among all evaluated models, with Dice scores of 0.8679 for LV, 0.8013 for RV, and 0.7191 for MYO. Although transformer components offer powerful global feature modeling, their integration into the U-Net framework without extensive optimization led to suboptimal results for this task. The relatively poor performance suggests that the model may have suffered from overfitting or insufficient training data for fine-grained medical image segmentation. As a result, the Transformer-Based U-Net was less effective in accurately segmenting cardiac structures. While not explicitly used in this model as shown in Figure 3.13, the traditional U-Net often includes skip connections to combine encoder and decoder features. Utilizing skip connections in the Trans-UN could improve the model performance. In contrast to CNN-based U-Nets, which can learn effectively even with limited data, transformers require more extensive datasets to learn spatial relationships from scratch. Since the transformer components in this model were trained from scratch without pretraining, the performance may have been hindered.

The Jaccard Coefficients obtained by each U-Net model for the LV, RV, and MYO regions in the test set are presented in Table 4.2. All six U-Net variants demonstrated moderate to strong segmentation capabilities across different cardiac structures, though with notable variations in performance across models and anatomical regions.

The results in Table 4.2 highlight several key observations:

- The **Original U-Net (O-UN)** performed consistently well, achieving strong Jaccard scores across all three regions, particularly for the LV (0.8161) and MYO (0.6835). Its performance closely aligns with that of more advanced variants, indicating that despite its simpler architecture, O-UN is capable of reliable segmentation. The relatively balanced performance across all structures makes it a strong baseline model.

TABLE 4.2: EVALUATION OF U-NET VARIANTS USING TEST SET JACCARD COEFFICIENTS

Model	LV	RV	MYO
Original U-Net (O-UN)	0.8161	0.6739	0.6835
Residual U-Net (Res-UN)	0.7603	0.6194	0.6165
Attention U-Net (Atn-UN)	0.8081	0.6629	0.6712
Feature Pyramid U-Net (FP-UN)	**0.8279**	0.6684	**0.6852**
Feedback Residual U-Net (Feed-Res-UN)	0.8115	**0.6777**	0.6832
Transformer-Based U-Net (Trans-UN)	0.6849	0.5224	0.5016

- The **Residual U-Net (Res-UN)** showed a decline in performance compared to O-UN, especially in the MYO region (0.6165), where it recorded the lowest score among the non-Transformer variants. Although residual connections can theoretically enhance feature propagation and learning depth, their implementation in this model appears to under-perform, possibly due to increased model complexity or overfitting. Despite this, it achieves a moderate score for RV (0.6194), but overall results suggest room for improvement.

- The **Attention U-Net (Atn-UN)** demonstrated competitive results in LV (0.8081) and MYO (0.6712) segmentation, outperforming Res-UN in all three regions. The attention mechanism likely improved the model's ability to focus on salient regions, although its performance in RV segmentation (0.6629) remains slightly behind other high-performing models. These results indicate that spatial attention contributes positively, though it may not be sufficient alone to fully capture complex cardiac boundaries.

- The **Feature Pyramid U-Net (FP-UN)** achieved the highest Jaccard scores for both LV (0.8279) and MYO (0.6852), confirming its superior capability in capturing diverse spatial features through multi-scale representations. This suggests that the feature pyramid mechanism allows the model to effectively integrate fine and coarse contextual information, which is especially advantageous in segmenting intricate structures like the myocardium. While its RV score (0.6684) is not the highest, its overall performance is consistently strong.

- The **Feedback Residual U-Net (Feed-Res-UN)** performed best in RV segmentation (0.6777), marginally surpassing all other models. This indicates the benefit of incorporating feedback loops in refining the segmentation of smaller or less defined structures. Its scores for LV (0.8115) and MYO (0.6832) are also among the top performers, suggesting that the feedback mechanism may enhance spatial refinement when effectively integrated. Nonetheless, the improvements over O-UN and FP-UN are relatively modest.

- The **Transformer-Based U-Net (Trans-UN)** recorded the lowest Jaccard coefficients across all three regions, with notably poor performance in RV (0.5224) and MYO (0.5016). These results highlight the limitations of Transformer-based models in small medical imaging datasets, where the lack of inductive biases and need for large-scale training data can severely impact effectiveness. The model's limited performance indicates that, without additional architectural improvements or access to larger training datasets, it may not be well-suited for detailed segmentation tasks such as cardiac MRI. This is largely due to the higher parameter count in transformer-based architectures compared to other U-Net variants, which increases the likelihood of overfitting when trained on smaller datasets.

Overall, the results from the Jaccard evaluation corroborate the Dice Score findings, emphasizing the robustness of multi-scale feature extraction in FP-UN and the challenges faced by Transformer-based models in limited-data settings. These insights underline the importance of architectural choices in achieving high-quality segmentation in complex medical imaging tasks.

4.3 Segmentation Results Analysis

This section analyzes the segmentation results of the U-Net variants for 4 different image samples.

The original image depicts a slice from a cardiac MRI scan, showcasing the heart's anatomical structures, including the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The gray-scale intensity variations highlight tissue contrast, enabling the distinction of these regions from the surrounding structures. This serves as the input for segmentation models.

The ground truth mask represents the manually annotated segmentation of the image, where the LV, RV, and MYO are distinctly labeled. Each region is assigned a specific intensity value to indicate its boundaries and areas accurately. The ground truth serves as a benchmark for evaluating the performance of segmentation models, ensuring their predictions align with expert annotations. It highlights the complexity of the task, especially in delineating thin and irregular regions like the myocardium.

In the first sample image (Figure 4.1), the Original U-Net captures the main structure of the LV, RV, and MYO, but minor discrepancies are visible along the borders. The segmentation is slightly blurred, with some overlap errors between regions (e.g., the boundary of the myocardium is not perfectly aligned with the ground truth). While effective as a baseline model, the Original U-Net struggles to capture sharp edges and fine details of the myocardium, leading to slight under-segmentation of the thin regions. This is consistent with its moderate Dice score for MYO.

The ResU-Net, on the other hand, demonstrates improved boundary alignment compared to the Original U-Net. The contours of the LV and RV are sharper and more

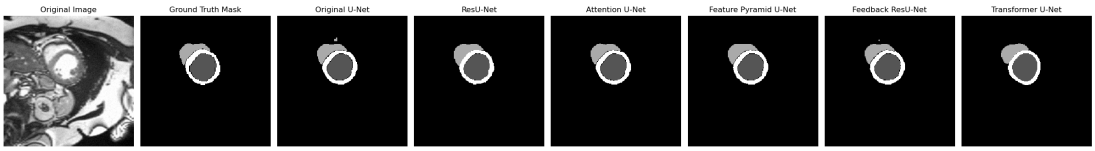

Fig. 4.1: Model Predictions for Sample Image - I

consistent with the ground truth, with fewer false positive regions. However, slight over-segmentation is noticeable at the boundaries of the myocardium. The residual connections appear to help the model capture the complex structures of the RV, as evidenced by its superior Dice score for this region. The segmentation for MYO shows modest improvement, but still lacks complete precision in thinner regions.

The Attention U-Net achieves a sharper focus on relevant regions, showing better boundary alignment for the LV and RV compared to the Original U-Net. However, the model appears to slightly under-segment the myocardium, leading to gaps or missing regions compared to the ground truth. The attention mechanism helps focus on larger structures like the LV, but it might struggle with fine-grained details, particularly for the MYO. The under-segmentation of MYO aligns with its slightly lower Dice score in this region.

The Feature Pyramid U-Net achieves excellent boundary alignment and segmentation across all regions. The model captures the fine details of the myocardium more effectively than the other variants, with minimal overlap errors or missing regions. The segmentation closely matches the ground truth, particularly for the thin myocardial structure. The use of multi-scale feature extraction allows FP-UN to perform best for LV and MYO, as evidenced by its Dice scores. It represents the most balanced and accurate segmentation among all models.

The Feedback ResU-Net produces segmentation similar to the Original U-Net, with visible blurring and slight misalignment at the boundaries of the LV and RV. The myocardium segmentation is incomplete, with some regions missing or under-segmented. The feedback mechanism does not significantly enhance feature learning, as the model struggles with finer details and thinner regions. Its performance is slightly better than the baseline for RV but does not surpass other advanced variants.

The Transformer-Based U-Net significantly underperforms, producing incomplete and irregular segmentation for all regions. Large portions of the myocardium are either missed or poorly segmented, and the boundaries of the LV and RV are less defined. This model's reliance on large datasets and global attention mechanisms seems to hinder its ability to generalize for small, detailed datasets like this cardiac MRI. Its poor performance is reflected in the much lower Dice scores across all regions.

Now let's analyze the second sample image in Figure 4.2.

The second sample image illustrates the segmentation outputs for various U-Net variants compared to the ground truth mask for a different slice of cardiac MRI. The

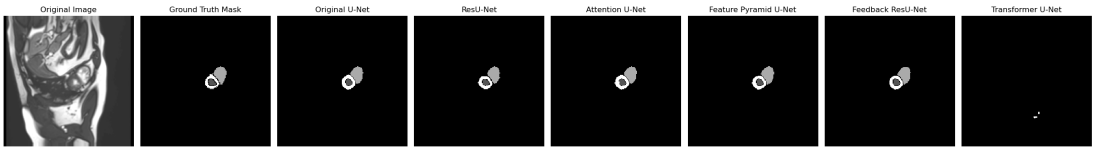

Fig. 4.2: Model Predictions for Sample Image - II

ground truth mask accurately delineates the left ventricle (LV), right ventricle (RV), and myocardium (MYO), with clear boundaries reflecting the regions of interest. The Original U-Net provides a reasonable segmentation but struggles with the precision of the myocardial boundary, showing slight overlap errors and boundary misalignment. The ResU-Net demonstrates improved performance with sharper contours, particularly for the LV and RV, but there is still minor over-segmentation around the myocardium. The Attention U-Net achieves well-defined LV and RV regions but under-segments the myocardium, failing to capture its full structure. The Feature Pyramid U-Net once again exhibits the most accurate segmentation, with clear boundary alignment and excellent representation of the myocardium, likely due to its multi-scale feature extraction capabilities. The Feedback ResU-Net produces a result similar to the Original U-Net, with noticeable blurring and slightly misaligned boundaries. Lastly, the Transformer-Based U-Net performs poorly, failing to segment the cardiac structures meaningfully, with only sparse, disjoint predictions visible. This result highlights the importance of dataset-specific adaptations and the need for inductive biases in Transformer-based approaches for small, detailed medical datasets. Overall, Feature Pyramid U-Net stands out as the most reliable model for this sample, excelling in its ability to capture both large and fine-grained details.

The third sample image, depicted in Figure 4.3, provides another comparison of U-Net variants' performance against the ground truth mask for cardiac MRI segmentation.

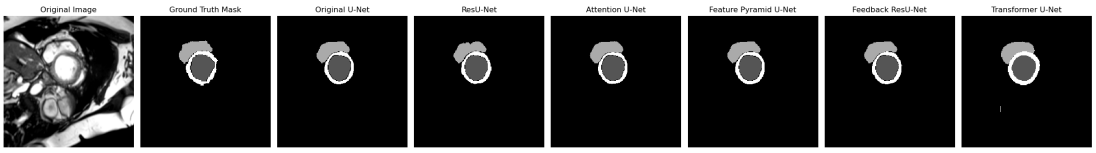

Fig. 4.3: Model Predictions for Sample Image - III

The ground truth mask clearly delineates the left ventricle (LV), right ventricle (RV), and myocardium (MYO), with distinct and well-defined boundaries. The Original U-Net produces reasonable segmentation, but struggles slightly with boundary precision, particularly for the RV, where some overlap errors are evident. The ResU-Net improves boundary sharpness, particularly for the LV, showing better alignment with the ground truth, though slight over-segmentation near the myocardium is visible. The Attention U-Net demonstrates sharper focus on the LV and RV regions but

under-segments the myocardium, leading to gaps in capturing its finer structure. The Feature Pyramid U-Net delivers the most accurate segmentation, with excellent boundary alignment and detailed myocardium segmentation, reflecting its ability to integrate multi-scale features effectively. The Feedback ResU-Net performs similarly to the Original U-Net, with noticeable blurring and minor misalignment in the RV and MYO regions, suggesting limited benefit from the feedback mechanism. The Transformer-Based U-Net shows slightly better performance compared to previous samples, successfully identifying the LV and RV, but its myocardium segmentation remains incomplete, with significant gaps and a lack of boundary precision. Overall, the Feature Pyramid U-Net outperforms other models in accurately segmenting all three cardiac regions, while the Transformer-Based U-Net remains limited in capturing fine-grained details.

The fourth sample image in Figure 4.4 compares the segmentation outputs of various U-Net variants for a cardiac MRI slice with the ground truth mask. Notably, the ground truth mask in this sample only delineates the left ventricle (LV) and myocardium (MYO), with no visible annotations for the right ventricle (RV), indicating its absence in the segmentation target.

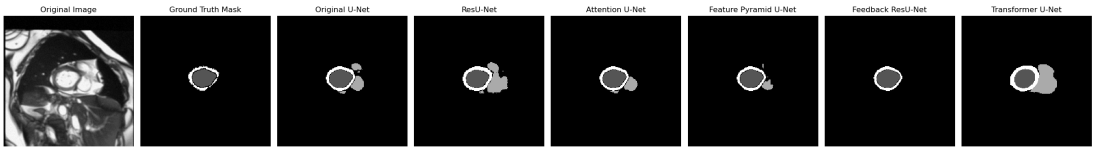

Fig. 4.4: Model Predictions for Sample Image - IV

The Original U-Net identifies the LV and MYO reasonably well but introduces false positives, especially in areas where the RV would typically be expected, leading to unnecessary segmentation artifacts. The ResU-Net slightly over-segments around the LV and MYO regions, producing artifacts that are not present in the ground truth, further complicating the evaluation. The Attention U-Net delivers sharper boundaries for the LV and MYO but under-segments some portions of the myocardium, failing to capture its full structure. The Feature Pyramid U-Net, despite its typically strong performance, introduces slight segmentation errors and artifacts, particularly along the myocardium boundaries, which deviate from the ground truth. The Feedback ResU-Net emerges as the best-performing model for this sample, providing smooth and precise segmentation of the LV and MYO with minimal false positives and a close alignment with the ground truth mask. The Transformer-Based U-Net under-performs, introducing significant over-segmentation artifacts, particularly in areas outside the ground truth-defined regions, highlighting its struggles with finer boundary precision in this specific case. Overall, Feedback ResU-Net demonstrates superior accuracy and reliability for this particular sample, effectively segmenting the LV and MYO without introducing unnecessary artifacts.

4.4 Dice Score and Loss Graphs

Figure 4.5 provides dice scores and losses of each U-Net variant trained in this study. Each subplot visualizes the training/ validation set dynamics of a specific U-Net variant across epochs. A summary of the metrics used in the plots are provided below.

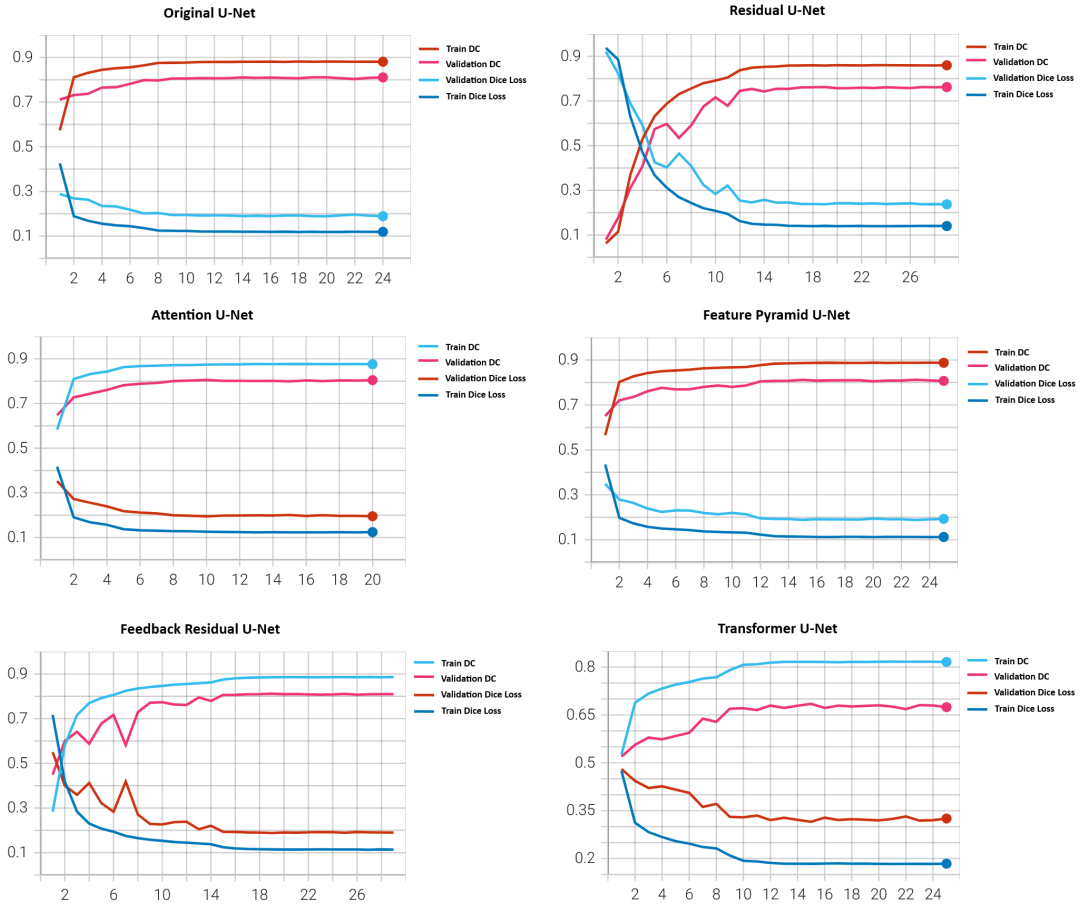

Fig. 4.5: Dice Score and Loss of U-Net Variants

- **Train DC (Dice Coefficient):** This metric quantifies the overlap between the predicted segmentation masks and the ground truth annotations within the training dataset. A higher Train DC (approaching 1) indicates better performance on the training data. The trend of Train DC reveals the model's learning progress during training. A rapid initial increase followed by a gradual plateau is typical, signifying effective learning in the early stages and diminishing returns as training progresses.

- **Validation Dice Coefficient (Validation DC):** This metric evaluates the model's performance on a held-out validation dataset, which is crucial for assessing generalization capability. A high Validation DC suggests that the model effectively

generalizes to unseen data. The relationship between Train DC and Validation DC is paramount. Ideally, the Validation DC should closely track the Train DC. A significant divergence between these two curves is a strong indicator of overfitting, where the model performs well on the training data but poorly on unseen data.

- **Training Dice Loss:** The Dice Loss is one of the aspects of the hybrid loss function minimized during training. It is inversely related to the Dice Coefficient. A decreasing Train Dice Loss signifies effective learning, as the model's predictions become increasingly similar to the ground truth. The rate of decrease in Train Dice Loss can also provide insights into the training dynamics.

- **Validation Dice Loss:** This metric measures the Dice Loss on the validation set. It serves as an independent evaluation of the model's performance on unseen data and is a key indicator of overfitting. If the Validation Dice Loss starts to increase after an initial decrease, it implies that the model is beginning to overfit the training data.

When observing the plots, we can see a rapid initial improvement in some of the U-Net variants with respect to (w.r.t) the Train DC. A steep initial rise in Train DC and a corresponding rapid decrease in Train Dice Loss indicate efficient learning at the beginning of training. Moreover, as the training progresses of the U-Net variants, the rate of improvement typically slows down, leading to a plateau in the Train DC and Train Dice Loss curves. This signifies that the models are approaching their learning capacity on the training data.

The term Overfitting (high Train DC, low Validation DC, increasing Validation Loss) refers to the significant gap between Train DC and Validation DC, coupled with an increasing Validation Dice Loss after an initial decrease. This means a model has memorized the training data but fails to generalize to unseen examples. While none of the plots in this study show severe overfitting (a clear increase in validation loss after an initial decrease), monitoring the gap between training and validation metrics will be important.

On the other hand, Underfitting refers to low values for both Train DC and Validation DC. This indicates that the model has not learned the underlying patterns in the data, possibly due to insufficient model capacity, inadequate training time, or suboptimal hyperparameters. However, none of the plots suggests that the U-Net variants trained in this study are prone to underfitting.

Now let's analyze the performance of each U-Net variant below.

- **Original U-Net (Baseline):** The original U-Net architecture serves as a baseline for our comparison. Its performance provides a reference point for evaluating the

effectiveness of the modifications introduced in the other variants. Its subplot shows reasonable training and validation Dice scores. The validation loss seems to stabilize, suggesting decent generalization.

- **Residual U-Net:** Appears to converge faster and potentially achieves slightly higher validation Dice scores than the original U-Net. The use of residual connections likely helps with optimization and performance.
- **Attention U-Net:** Seems to perform similarly to the Residual U-Net, possibly with a slightly quicker initial improvement. Attention mechanisms help the network focus on relevant features, which can improve segmentation accuracy.
- **Feature Pyramid U-Net:** This architecture also performs well, achieving high Dice scores. They are designed to capture multi-scale information, which is beneficial for segmenting objects of varying sizes.
- **Feedback Residual U-Net:** This variant seems to have some instability in the initial training phase, but eventually reaches competitive performance. The feedback connections may introduce some training challenges.
- **Transformer U-Net:** This architecture shows slower convergence and potentially slightly lower performance compared to other variants. Transformers, while powerful, can require more data and careful tuning for image segmentation tasks.

When comparing the subplots, the Residual U-Net exhibited faster convergence compared to the original U-Net, achieving a higher Validation DC with fewer training epochs. This suggests that the residual connections facilitated more efficient training. The Attention U-Net further improved performance, likely due to its ability to focus on relevant image features. The Feature Pyramid U-Net architecture also demonstrated strong performance, highlighting the importance of multi-scale feature representation for this segmentation task. While the Feedback Residual U-Net showed promise, its training was less stable. The Transformer U-Net exhibited slower convergence, suggesting the need for more extensive training data or pre-training.

In this study, training was conducted for a maximum of 100 epochs, with early stopping implemented to halt training when the validation performance ceased to improve, reducing the risk of overfitting. This strategy is evident in the plots in Figure 4.5, where the training curves for all six U-Net variants exhibit varying degrees of stabilization. For instance, the Residual U-Net demonstrates relatively rapid convergence, with both training and validation metrics reaching a plateau within the first 20 epochs, suggesting that further training would likely yield minimal gains. In contrast, the Transformer U-Net exhibits a more gradual learning curve, requiring a larger number of

epochs to reach a stable state. Early stopping prevented the models from overfitting the training data, ensuring that they maintained good generalization capabilities and performed well on unseen data.

The subplots in Figure 4.6 depicts the training hybrid loss for each U-Net variant. These curves represent the combined effect of the cross-entropy and Dice loss components, providing a comprehensive view of the training progress. A rapid initial decrease in the loss indicates effective learning in the early stages, while a gradual plateau suggests that the model is approaching convergence.

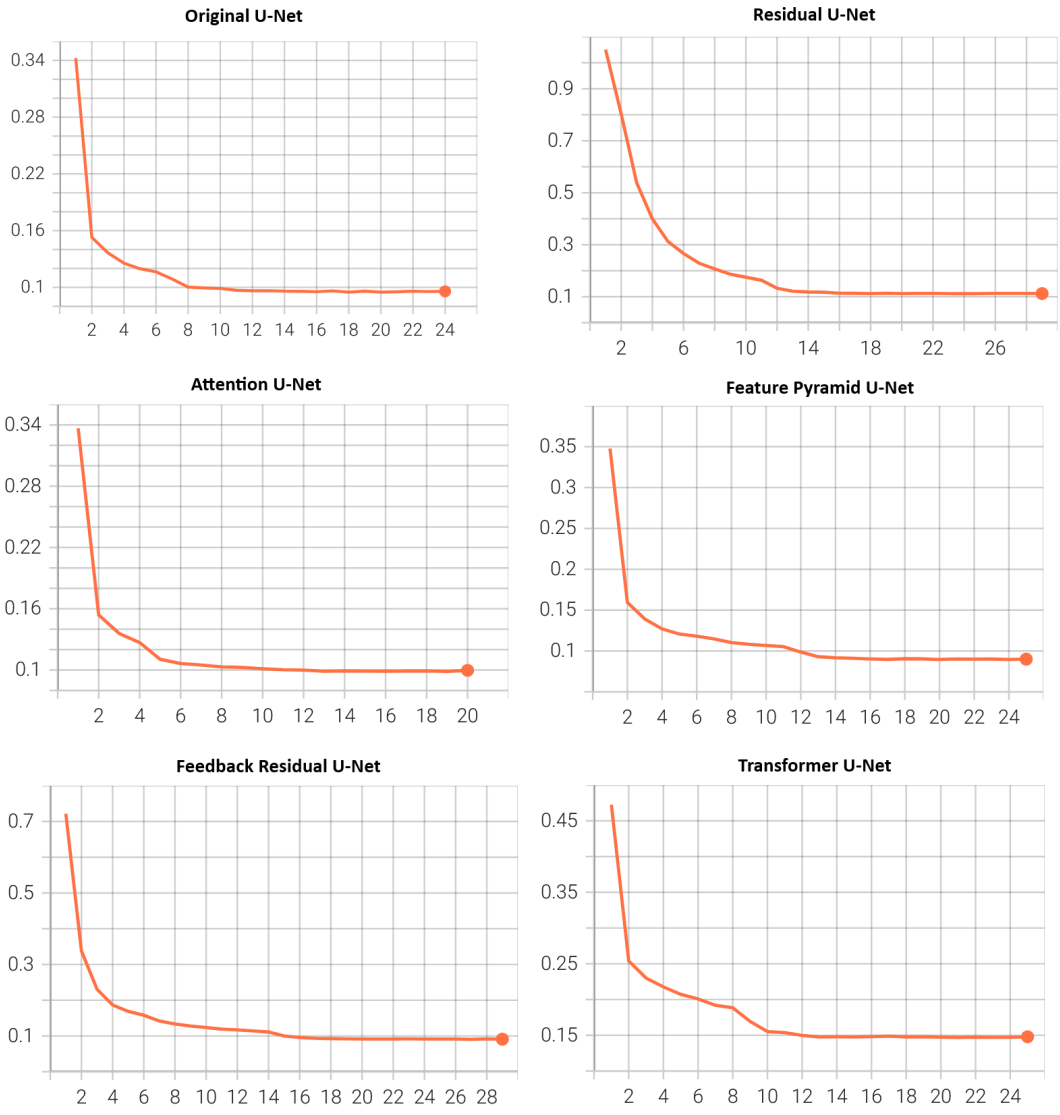

Fig. 4.6: Training Hybrid Loss of U-Net Variants

The hybrid loss for the Original U-Net shows a steep decrease during the initial few epochs, converging at around 0.1 after approximately 12 epochs. This pattern indicates that the basic U-Net architecture effectively optimizes the combined loss, but reaches a performance plateau relatively early compared to more advanced architectures.

The Residual U-Net demonstrates a rapid initial reduction in hybrid loss, with a convergence point similar to the Original U-Net at around 0.1. However, it starts with a significantly higher initial loss (near 0.9), suggesting that while it eventually performs comparably, the initial optimization process might be slower or less stable.

The Attention U-Net exhibits a hybrid loss curve with an initial value similar to the Original U-Net but converges more smoothly and slightly earlier. The attention mechanism incorporated into the model likely aids in more targeted feature learning, leading to slightly more efficient convergence.

The hybrid loss for the Feature Pyramid U-Net starts at a higher initial value (around 0.35) and decreases rapidly, stabilizing at approximately 0.1 after around 20 epochs. The use of multi-scale feature extraction in this architecture appears to enhance early-stage learning, though the final convergence point is comparable to other variants.

The Feedback Residual U-Net presents a unique behavior with an initial hybrid loss of about 0.7, followed by a rapid decline and stabilization around 0.1 after nearly 25 epochs. The feedback mechanism might contribute to slightly delayed convergence, but aids in achieving comparable final performance.

The Transformer U-Net begins with a higher initial loss near 0.45, converging at approximately 0.15 after about 20 epochs. Unlike other models, it does not reach the same low hybrid loss level as the others, which could be attributed to the different optimization dynamics introduced by the self-attention mechanism in the transformer architecture. This could also indicate that the transformer architecture requires more training data to effectively segment the anatomically significant regions in the Cardiac MRI images. This is often observed with Transformer-based models, especially when trained on relatively smaller datasets.

The hybrid loss curves reveal that while all models ultimately converge to similar loss levels (between 0.1 and 0.15), there are notable differences in their initial behavior and convergence rates. The advanced U-Net variants (Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer U-Net) tend to start with higher initial hybrid losses compared to the Original U-Net, indicating the added complexity of these architectures.

Interestingly, although the Transformer U-Net shows a slower overall convergence and a slightly higher final hybrid loss, its ability to capture long-range dependencies may provide other benefits, such as improved generalization, which might not be reflected purely in loss values. Similarly, the Feedback Residual U-Net takes longer to converge fully but may benefit from its iterative refinement mechanism.

The common convergence point across most models suggests that, in terms of hybrid loss, there may be diminishing returns in adding architectural complexity beyond a certain point. However, further evaluation on validation or test sets could provide deeper insights into the generalization capabilities of each model.

CHAPTER 5

DISCUSSION

Cardiac MRI segmentation is a critical task in medical imaging, enabling precise delineation of cardiac structures for diagnostic and therapeutic purposes. In this research, we aimed to evaluate and develop U-Net-based architectures to enhance the segmentation accuracy of cardiac MRI data. The study encompasses the implementation and analysis of six U-Net variants, namely Original U-Net, Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer-Based U-Net, leveraging their unique capabilities to tackle the challenges posed by cardiac MRI datasets, such as variability in anatomy, low contrast, and noise. Each model is trained and evaluated rigorously to quantify its performance and identify its strengths and limitations.

The results demonstrate that the Feature Pyramid U-Net achieved the best overall performance across the three structures, with Dice coefficients of 0.9388 (LV), 0.8759 (RV), and 0.8426 (MYO), showcasing its ability to leverage multi-scale feature extraction for precise segmentation. The ResU-Net model performed particularly well for RV segmentation, achieving the highest score among variants (0.8812), highlighting the effectiveness of residual connections to capture complex features.

In order to minimize overfitting and improve model generalization, several strategies were employed during training. These included the use of early stopping, which stopped training when validation performance ceased to improve, preventing overfitting to training data. The AdamW optimizer, with its decoupled weight decay, further improved generalization by regularizing the model weights. These measures collectively ensured robust performance across the test set.

5.1 Study Contribution

This research makes significant contributions to the field of medical image segmentation, with a particular focus on the segmentation of cardiac MRI images. By developing and analyzing six distinct U-Net variants: Original U-Net, Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer-Based U-Net, this study advances the understanding of how architectural innovations impact the accuracy and robustness of segmentation models in this critical domain. Below are the key contributions of this study:

1. **Development of Novel Architectures:** The study introduces five novel U-Net variants: the Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net and the Transformer-Based U-Net specifically tailored for

cardiac MRI segmentation. Unlike existing models, the Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback ResU-Net and Transformer-Based U-Net are refined and adapted as part of this work to better address the challenges of cardiac MRI segmentation, highlighting the contributions made in enhancing these architectures.

2. **Development of a Hybrid Loss Function:** One of the unique contributions of this study is the design and application of a hybrid loss function tailored for cardiac MRI segmentation. Segmentation tasks often suffer from imbalanced class distributions, where the regions of interest (e.g., ventricles or myocardium) occupy a small portion of the image compared to the background. To mitigate this issue, the proposed novel hybrid loss function combines Dice Loss and Cross-Entropy Loss, leveraging the strengths of both. The *Dice Loss* addresses class imbalance by directly optimizing the overlap between the predicted and ground truth segmentation, ensuring precise boundary delineation. On the other hand, the *Cross-Entropy Loss* enhances the network's ability to learn pixel-wise classification, ensuring that both large and small regions are segmented accurately. This hybrid approach balances global overlap optimization with local pixel-level accuracy, contributing to superior performance across different cardiac structures. By integrating this loss function into all U-Net variants, the study demonstrates its effectiveness in improving segmentation outcomes, particularly in scenarios involving low-contrast or irregularly shaped structures.

3. **Development of a Web Application:** One of the key contributions of this research is the development and deployment of an interactive web application for cardiac MRI segmentation using the ACDC dataset. Built with a Feature Pyramid UNet backbone and hosted on Hugging Face Spaces, the application provides a streamlined interface for uploading images, viewing segmented anatomical structures, and receiving contextual descriptions of selected regions. This tool bridges the gap between complex deep learning models and clinical usability by eliminating the need for technical expertise in model execution. The application demonstrates how AI-powered segmentation can be translated into accessible, real-time decision support tools, reinforcing its potential integration into clinical workflows and educational settings.

4. **Comprehensive Evaluation of U-Net Variants:** One of the primary contributions of this research is the systematic evaluation of U-Net variants in the context of cardiac MRI segmentation. Each variant incorporates unique architectural enhancements that address specific limitations of traditional convolutional networks. This comparative analysis offers a deeper understanding of how these innovations, such as residual learning, attention mechanisms, feature pyra-

mids, feedback loops, and transformers, affect the segmentation performance across different cardiac structures. The findings provide valuable insights for researchers and practitioners seeking to select or design models tailored to their specific use cases.

5. **Impact of Architectural Enhancements:** The study highlights the impact of specific architectural components on segmentation outcomes. For example, it demonstrates how attention mechanisms improve the localization of small or low-contrast cardiac structures, how residual connections enhance gradient flow for deeper networks, and how transformers provide a global context that complements local feature extraction. These insights contribute to the broader understanding of deep learning in medical imaging.

6. **Practical Implications for Clinical Use:** By improving the segmentation accuracy and robustness of cardiac MRI, this research has direct implications for clinical applications. Accurate segmentation is a prerequisite for tasks such as quantifying cardiac volumes, assessing myocardial function, and detecting abnormalities. The proposed models, such as the Feature Pyramid U-Net, Feedback Residual U-Net and the Transformer-Based U-Net, have the potential to reduce manual intervention, streamline workflows, and enhance the reliability of cardiac MRI analysis in clinical settings.

7. **Foundation for Future Research:** The findings of this study lay a strong foundation for future research in cardiac imaging and beyond. The architectural innovations explored here can be further extended and applied to other medical imaging modalities, such as CT or ultrasound. Additionally, the incorporation of advanced techniques, such as data augmentation, domain adaptation, and multimodal fusion, could build upon the contributions of this research to push the boundaries of medical image segmentation.

8. **Model Extensibility and Future Adaptability:** One of the key strengths of the developed U-Net variants, especially the Feature Pyramid U-Net, Feedback Residual U-Net and Transformer-based U-Net is their modular and extensible architecture. This makes it feasible to adapt or enhance the models without rebuilding them from scratch. Thus, future extensions can involve replacing or tuning optimizers (e.g. trying Lookahead optimizers), refining or combining loss functions such as incorporating focal loss or Tversky loss, applying self-supervised learning or semi-supervised learning to reduce dependency on labeled data, and combine proposed U-Net variants through ensemble methods to further enhance the CMRI segmentation performance of ventricular structures and myocardium.

In summary, this research provides valuable insights into the design and application of deep learning models for medical imaging. By bridging the gap between architectural innovation and practical utility, this study contributes to the development of more accurate, efficient, and clinically relevant segmentation solutions.

5.1.1 Achieving Research Objectives and Research Questions

This study aimed to address key challenges in cardiac MRI (CMRI) segmentation through the development and evaluation of U-Net-based architectures. The research objectives, outlined at the beginning of this study, were systematically achieved through a detailed and structured approach. The following discussion highlights how each objective and corresponding research question was addressed.

5.1.1.1 Propose U-Net-based Architectures for Accurate CMRI Segmentation

To achieve this objective, six U-Net variants were designed and evaluated: Original U-Net, Residual U-Net, Attention U-Net, Feature Pyramid U-Net, Feedback Residual U-Net, and Transformer-Based U-Net. These architectures were selected and refined to address specific challenges in cardiac MRI segmentation, such as handling imbalanced datasets, capturing fine-grained details, and incorporating global context.

Unlike existing studies that often propose a single, highly customized architecture, this research conducted a **systematic comparative analysis** of multiple U-Net variants. This approach not only enabled a thorough understanding of their relative strengths and weaknesses but also ensured a more versatile and comprehensive solution. The proposed architectures, while simpler than many state-of-the-art methods, demonstrated sufficient accuracy and robustness for practical applications, making them accessible and reproducible across different datasets and clinical settings.

5.1.1.2 Enhance the Accuracy in Delineating the Boundaries of Cardiac Structures

The study focused on improving boundary delineation, which is critical for accurate segmentation of the ventricles and myocardium. This was achieved by integrating architectural enhancements tailored to address specific issues:

- **Residual U-Net** introduced residual connections to enhance gradient flow, enabling deeper networks to capture more detailed features.
- **Attention U-Net** employed attention mechanisms to focus on the regions of interest, suppressing background noise and improving boundary localization.
- **Feedback Residual U-Net** incorporated iterative refinement processes to correct segmentation errors progressively, leading to more precise boundary delineation.

Additionally, the use of a **hybrid loss function** combining Dice Loss and Cross-Entropy Loss contributed significantly to improving segmentation accuracy, particularly for small or low-contrast regions.

5.1.1.3 Compare the Performance with Existing Methods

A detailed comparative analysis was conducted to benchmark the proposed architectures against existing methods for CMRI segmentation. Dice Similarity Coefficient (DSC), a widely used performance evaluation metric in medical image segmentation tasks, was used to quantify the effectiveness of each U-Net variant.

The results demonstrated that the proposed architectures, particularly the Residual U-Net, Feature Pyramid U-Net, and Feedback Residual U-Net, achieved comparable accuracy to state-of-the-art methods while maintaining lower computational complexity. This balance of performance and efficiency makes the proposed methods practical for widespread adoption in both research and clinical applications.

The systematic evaluation also provided valuable insights into the strengths and limitations of each variant, enabling researchers and practitioners to select the most appropriate architecture based on specific requirements, such as computational resources or dataset characteristics.

5.1.1.4 Utilize the CMRI Segmentation Model for Inferencing in Clinical Settings

This study developed a lightweight, browser-accessible web application that seamlessly integrates deep learning-based segmentation into a clinical workflow. By deploying the best performing Feature Pyramid U-Net model through a Gradio interface hosted on Hugging Face Spaces, the application allows users, including clinicians, radiologists, and medical students, to perform inference on cardiac MRI images without requiring specialized hardware or software installations.

The app supports intuitive image upload, real-time segmentation visualization, class-wise color-coded annotations, and detailed anatomical descriptions upon selection, enhancing both interpretability and usability. This implementation demonstrates a practical pathway for translating research-grade models into clinically relevant tools, ensuring accessibility, ease of use, and rapid integration within diagnostic or educational environments.

5.2 Comparison with Existing Studies

Table 5.1 provides a comparison of the Dice Coefficients of the proposed U-Net variants with existing state-of-the-art methods for CMRI segmentation, conducted using the ACDC dataset. When considering LV segmentation, the Feature Pyramid U-Net

achieved a Dice score of 0.945, which is competitive with several state-of-the-art methods, but falls short of the top-performing Sharan *et al.* [15] with a Dice score of 0.958. For MYO segmentation, the Feature Pyramid U-Net achieved the highest score among the tested variants at 0.851, outperforming the other four variants. However, compared to existing studies, it underperformed by Sharan *et al.* [15] and Kamal *et al.* [14], which achieved a Dice score of 0.914. The Original UNet achieved the highest score among the variants for RV segmentation at 0.891, surpassing Feature Pyramid U-Net (0.888). However, both models lag behind Sharan *et al.* [15], which achieved a score of 0.934.

Although studies like [4, 14, 15] have reported high Dice scores, one limitation is that their evaluations were performed using only the ACDC training set or a small subset thereof, without incorporating the official ACDC test set as stated in Table 5.2. This can introduce a bias toward the training data and limits the assessment of true generalizability. In contrast, our study uniquely evaluates the models on the official ACDC test set, ensuring an unbiased and realistic evaluation that aligns with the standards of the ACDC 2017 Challenge.

Another distinguishing strength of our approach is the development and evaluation of six U-Net variants, incorporating architectural innovations such as residual connections, attention mechanisms, feature pyramids, feedback loops, and transformer modules. These enhancements improve multi-scale feature extraction, long-range dependency modeling, and fine-grained segmentation, addressing shortcomings in prior works which largely relied on standard encoder-decoder structures without such advanced features. While real-time inference capability (2 seconds per image) is another advantage of our method (refer Table 5.2), the scientific contribution goes beyond speed: our model's architectural improvements lead to a balanced optimization of accuracy, generalizability, and computational efficiency, all critical for clinical deployment.

A particularly noteworthy aspect of our study is that these competitive results were achieved despite operating with relatively modest computational resources as highlighted in Table 5.2. In contrast to many state-of-the-art models (e.g., [14], [49]), which leveraged high-end hardware such as the NVIDIA Quadro RTX 5000 GPU or the powerful A100 Tensor Core GPUs, our models were trained using a single NVIDIA RTX 2080 Super GPU. Despite fully utilizing the available memory capacity, we did not have access to large-scale parallelization or specialized hardware accelerators, which are often used to further boost training efficiency and model performance. As a result of these hardware constraints, we were limited to smaller batch sizes during training. It is well-known that larger batch sizes can contribute to more stable gradient updates, potentially leading to faster convergence and improved final model performance. The fact that our models still achieved strong results under these constraints highlights the robustness, efficiency, and practicality of our solutions, qualities that are crucial for

TABLE 5.1: DICE SCORE COMPARISON WITH EXISTING STUDIES EVALUATED ON ACDC DATASET

Study	Method	LV	RV	MYO
[12]	SAUNet - Shape Attentive U-Net	0.938	0.914	0.887
[5]	Res U-Net as the initial segmentation network and a hierarchical ConvLSTM based recurrent network as the temporal consistency network	0.855	0.761	0.746
[4]	Cascaded approach: UNet for ROI extraction, FCN for initial segmentation, and a U-Net model for refinement	0.938	0.880	0.900
[9]	Multi-Task Learning based U-Net (MTLUNET)	0.881	0.724	0.807
[15]	U-Net with VGG encoder and Feature Pyramid Network	0.958	0.934	0.914
[14]	Attention-guided Residual W-Net (ARW-Net)	0.953	0.923	0.914
[49]	UU-Mamba model: integrating the U-Mamba model with the Sharpness-Aware Minimization (SAM) optimizer and an uncertainty-aware loss function	0.950	0.924	0.909
Ours	Original U-Net	0.941	0.891	0.851
Ours	ResU-Net	0.922	0.845	0.820
Ours	Attention U-Net	0.937	0.881	0.840
Ours	Feature Pyramid U-Net	0.945	0.888	0.851
Ours	Feedback ResU-Net	0.939	0.890	0.846
Ours	Transformer-Based U-Net	0.868	0.801	0.719

LV: Left Ventricle, RV: Right Ventricle, MYO: Myocardium.

clinical translation, especially in settings where advanced computational infrastructure is not readily available.

Furthermore, a key differentiating factor of our study is its **support for real-time prediction** using a web application, which is absent in all the other studies. Real-time capability is crucial for clinical integration, particularly in time-sensitive environments where rapid and accurate segmentation can significantly impact diagnosis and treatment decisions.

In conclusion, our study provides a comprehensive evaluation of six enhanced U-Net variants for cardiac MRI segmentation, addressing key gaps in existing literature. Unlike many prior works that evaluated models only on the ACDC training set, our approach ensures unbiased assessment using the official test set. Despite using modest computational resources, our models achieved competitive performance, highlighting their efficiency and robustness. Architectural enhancements and real-time prediction

TABLE 5.2: COMPARISON WITH EXISTING STUDIES EVALUATED ON ACDC DATASET

Study	**# Training Set Subjects**	**# Validation Set Subjects**	**# Test Set Subjects**	**Inference Time**	**Remarks**
[12]	80	50	20	Not specified	Trained using NVIDIA Quadro RTX 5000 GPU with 16GB of memory
[5]	70	20	10	Not specified	Trained with limited GPU memory
[4]	60	20	20	4 s per patient	Trained using NVIDIA GTX Titan X or V high-performance GPUs
[9]	80	Not specified	20	Not specified	Trained using 8GB NVIDIA GTX 1080 GPU
[15]	60	20	20	Not specified	Trained using NVIDIA K80 GPU provided by Google Collaboratory
[14]	100	Not specified	50	Not specified	Trained using NVIDIA Quadro RTX 5000 for 1000 epochs
[49]	60	20	20	Not specified	Trained using two NVIDIA A100 Tensor Core GPUs
Ours	120	15	15	2 s per image	Trained using NVIDIA GeForce RTX 2080 Super GPU

capability further distinguish our work, making it well-suited for practical clinical deployment.

5.3 Challenges and Limitations

Despite the significant contributions of this research to cardiac MRI segmentation, the study encountered several challenges and limitations, primarily related to compu-

tational resources, platform constraints, and infrastructure issues. These limitations provide valuable insights into the practical aspects of conducting machine learning research and highlight areas for improvement in future studies.

5.3.1 Computational Resource Limitations

Training deep learning models for medical image segmentation is computationally intensive, requiring substantial GPU memory and processing power. This study faced several challenges due to limited resources:

- **GPU Constraints:** The models were initially trained on a GPU laptop, which significantly prolonged the training time. Some models, such as the Transformer-Based U-Net and Feature Pyramid U-Net, required more than 24 hours for a single training session due to their complexity and high memory demands.
- **GPU Memory Limitations:** The GPU often ran out of memory when processing larger batch sizes or training deeper models. This required reducing batch sizes and optimizing memory usage, which sometimes led to slower convergence or suboptimal performance.

5.3.2 Platform Constraints

The use of cloud-based platforms like Google Colab presented additional challenges:

- **Session Timeouts:** The free version of Google Colab imposes a maximum session duration of 12 hours, which was insufficient for training more complex architectures. Interrupted training sessions required restarting from the last checkpoint, leading to inefficiencies and delays.
- **High Cost of Premium Services:** To overcome the limitations of the free version, the Google Colab Premium version was used on one occasion. However, the high subscription cost made it impractical for sustained use throughout the study, limiting access to more powerful computational resources.

5.3.3 Changes in External Infrastructure

In this study, for tracking the experiments, including model management, and tracking performance metrics, MLflow was used. The free MLflow service provided by the Databricks Community Edition was used in the study. However, the basic authentication feature was removed during the project, rendering the setup unusable. As a result, the MLFlow tracking system had to be migrated and set up locally. This unplanned migration consumed valuable time and effort, diverting attention from core research tasks.

5.3.4 Challenges in Model Complexity and Training

The complexity of the proposed architectures, particularly the Feedback Residual U-Net and Transformer-Based U-Net, introduced unique challenges:

- **Long Training Times:** These models required extended training durations to achieve convergence, a challenge compounded by the constraints of limited computational resources.
- **Hyperparameter Tuning:** Fine-tuning hyperparameters for each U-Net variant was computationally expensive and time-consuming, particularly for architectures involving iterative refinement or attention mechanisms.

In summary, the challenges encountered during this study underline the critical importance of adequate computational resources and stable infrastructure in conducting machine learning research. Limited GPU memory, prolonged training times, and interruptions from platform constraints were significant hurdles that affected the efficiency of the research process. Changes in external tools, such as MLFlow on Databricks, further complicated the workflow, necessitating alternative solutions.

These limitations, while challenging, also provided valuable learning experiences that informed the research process. Future work can address these issues by leveraging high-performance computing clusters, exploring cost-effective cloud solutions, and ensuring robust infrastructure setups to mitigate disruptions. Overcoming these limitations will pave the way for more efficient and scalable research in medical image segmentation.

5.4 Future Work

This research lays the groundwork for several promising directions that can be pursued to further advance the field of cardiac MRI segmentation. While the study has made significant strides in proposing U-Net-based architectures, several areas remain ripe for exploration, including model refinement, the incorporation of diverse imaging modalities, and enhancing computational efficiency. Future work will focus on addressing these aspects to improve segmentation accuracy, generalization capability, and clinical applicability.

- **Ensemble Learning Approaches:** Although the study systematically compared six U-Net variants, combining these models through ensemble learning could further enhance segmentation performance. Ensemble methods, which aggregate predictions from multiple models, have proven effective in reducing overfitting and improving generalization across diverse datasets. By leveraging the

strengths of different U-Net architectures, future research could explore ensemble techniques, such as bagging, boosting, or stacking, to create more robust and accurate segmentation models.

- **Incorporation of Additional Imaging Modalities:** Cardiac MRI provides detailed soft-tissue contrast, but its effectiveness can be complemented by other imaging modalities. Future research could explore the integration of CT (Computed Tomography) or echocardiography images, which offer complementary information on cardiac anatomy. Multi-modal learning approaches could improve the network's ability to generalize to different anatomical variations and clinical scenarios, enhancing segmentation accuracy across diverse patient populations.

- **Exploration of Feedback U-Net Variants:** Although the Feedback Residual U-Net showed promise, its performance was limited in this study due to computational constraints and fewer iterations. The iterative refinement mechanism holds significant potential for improving boundary accuracy, particularly in regions with ambiguous structures. Future work could explore deeper feedback mechanisms, enabling more sophisticated refinements across multiple iterations. This would allow the network to progressively learn from its mistakes and refine segmentation outcomes more accurately.

- **Development of Lightweight, Efficient Architectures:** With real-time clinical applications in mind, there is a growing need for computationally efficient and lightweight architectures that can be deployed in resource-limited clinical environments. Future work could focus on reducing model complexity while maintaining high segmentation accuracy. These lightweight models should also incorporate explainability techniques to increase trust in clinical decision-making. Such architectures would be valuable for applications where timely, interpretable results are critical, without sacrificing performance.

- **Application to Real-World Clinical Scenarios:** Expanding beyond the current publicly available datasets, future research should focus on applying these models to real-world clinical datasets that encompass a broader range of patient conditions, imaging quality, and anatomical variability. Real-world validation will provide a more accurate assessment of how well these U-Net variants can generalize to clinical practice, ensuring that the models are robust and reliable for practical use.

- **Use of Transfer Learning with Pre-Trained Models:** Transfer learning has emerged as a powerful technique in deep learning, where knowledge from large-scale models is transferred to smaller, domain-specific tasks. Future work could

explore pre-training models on larger cardiac datasets or general medical imaging datasets and fine-tuning these models specifically for CMRI segmentation. This approach would allow leveraging vast amounts of publicly available data, potentially improving generalization to unseen patient cases. Moreover, utilizing pre-trained models provides a substantial advantage over training models from scratch, as these models leverage knowledge learned from vast, general datasets and have the ability to rapidly converge to optimal solutions with reduced resource requirements.

- **Enhanced Model Explainability:** Improved model interpretability will be crucial for clinical adoption. Future work could focus on developing explainable AI (XAI) techniques, such as saliency maps, activation mapping, or attention visualization, that allow clinicians to better understand model decisions. This will increase trust in the model's predictions and facilitate its integration into clinical workflows.

CHAPTER 6

CONCLUSION

This study presented a comprehensive evaluation of various U-Net-based architectures for cardiac MRI segmentation, addressing a critical task in the diagnosis and management of cardiovascular diseases. By systematically comparing six U-Net variants, each incorporating architectural innovations such as attention mechanisms, feature pyramids, and transformers, this study provided a deeper understanding of their effectiveness in segmenting the left ventricle, right ventricle, and myocardium from short-axis Cardiac MRI slices. The Feature Pyramid U-Net emerged as the most robust model, achieving the highest Dice coefficients across all anatomical classes, thus demonstrating its strong capability to capture multi-scale contextual information and intricate structural boundaries.

In addition to performance benchmarking, this work emphasized practicality and accessibility by translating these deep learning models into a deployable format through a user-friendly web application. The web interface, hosted on Hugging Face Spaces, allows clinicians, researchers, and students to perform real-time inference on CMRI slices and interpret segmentation results with ease. This component shows how AI-driven solutions can be integrated into clinical workflows to support better diagnoses and make it easier for healthcare professionals to use machine learning tools.

The simplicity, reproducibility, and thoroughness of the experimental framework not only strengthen the credibility of the findings but also offer a solid foundation for future research in medical image segmentation. The insights drawn from this comparative study highlight both the strengths and limitations of existing U-Net variants, paving the way for the development of next-generation segmentation architectures. Furthermore, by addressing the feasibility of deploying these models in clinical settings, the research bridges the gap between academic exploration and practical application, ensuring its relevance in both theoretical and operational contexts.

Ultimately, this thesis contributes valuable knowledge to the evolving field of medical image analysis, emphasizing the importance of usability, deployment feasibility, and segmentation performance in real-world scenarios. The lessons learned and tools developed, particularly the comparative evaluation of U-Net variants and the deployment of a clinical-facing web application, are expected to guide future innovations and inspire continued advancements in deep learning for cardiac imaging.

REFERENCES

[1] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep Learning for Cardiac Image Segmentation: A Review," *Frontiers in Cardiovascular Medicine*, vol. 7, 2020.

[2] Y.-L. Lu, K. A. Connelly, A. J. Dick, G. A. Wright, and P. E. Radau, "Automatic functional analysis of left ventricle in cardiac cine MRI," *Quantitative imaging in medicine and surgery*, vol. 3, no. 4, p. 200, 2013.

[3] M.-P. Jolly, "Automatic segmentation of the left ventricle in cardiac MR and CT images," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 151–163, 2006.

[4] I. F. S. da Silva, A. C. Silva, A. C. de Paiva, and M. Gattass, "A cascade approach for automatic segmentation of cardiac structures in short-axis cine-MR images using deep neural networks," *Expert Systems with Applications*, vol. 197, p. 116704, 2022.

[5] Y. Chen, W. Xie, J. Zhang, H. Qiu, D. Zeng, Y. Shi, H. Yuan, J. Zhuang, Q. Jia, Y. Zhang *et al.*, "Myocardial segmentation of cardiac MRI sequences with temporal consistency for coronary artery disease diagnosis," *Frontiers in Cardiovascular Medicine*, vol. 9, p. 804442, 2022.

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

[7] E. Shibuya and K. Hotta, "Feedback U-Net for cell image segmentation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, 2020, pp. 974–975.

[8] B. Wu, Y. Fang, and X. Lai, "Left ventricle automatic segmentation in cardiac MRI using a combined CNN and U-net approach," *Computerized Medical Imaging and Graphics*, vol. 82, p. 101719, 2020.

[9] J. Ren, H. Sun, H. Zhao, H. Gao, C. Maclellan, S. Zhao, and X. Luo, "Effective extraction of ventricles and myocardium objects from cardiac magnetic resonance images with a multi-task learning U-Net," *Pattern Recognition Letters*, vol. 155, pp. 165–170, 2022.

[10] M. A. Al-antari, Z. Farea Shaaf, M. Mahadi Abdul Jamil, N. Abdel Samee, R. Alkanhel, M. Talo, and Z. Al-Huda, "Deep learning myocardial infarction segmentation framework from cardiac magnetic resonance images," *Biomedical Signal Processing and Control*, vol. 89, p. 105710, 2024.

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

[12] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "Saunet: Shape attentive u-net for interpretable medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 797–806.

[13] C. Hengfei, Y. Chang, J. Lei, X. Yong, and Z. Yanning, "Multiscale attention guided U-Net architecture for cardiac segmentation in short-axis MRI images," *Computer Methods and Programs in Biomedicine*, 2021.

[14] K. Raj Singh, A. Sharma, and G. Kumar Singh, "Attention-guided residual W-Net for supervised cardiac magnetic resonance imaging segmentation," *Biomedical Signal Processing and Control*, vol. 86, p. 105177, 2023.

[15] T. S. Sharan, S. Tripathi, S. Sharma, and N. Sharma, "Encoder Modified U-Net and Feature Pyramid Network for Multi-class Segmentation of Cardiac Magnetic Resonance Images," *IETE Technical Review*, vol. 39, no. 5, pp. 1092–1104, 2022.

[16] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *CoRR*, vol. abs/1711.10684, 2017. [Online]. Available: http://arxiv.org/abs/1711.10684

[17] W. Y. I. Tseng, M. Y. M. Su, and Y. H. E. Tseng, "Introduction to Cardiovascular Magnetic Resonance: Technical Principles and Clinical Applications," *Acta Cardiologica Sinica*, 2016.

[18] O. Bernard *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[19] A. Ammari, R. Mahmoudi, B. Hmida, R. Saouli, and M. H. Bedoui, "A review of approaches investigated for right ventricular segmentation using short-axis cardiac MRI," *IET Image Processing*, vol. 15, no. 9, pp. 1845–1868, 2021.

[20] C. Petitjean and J. N. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Medical Image Analysis*, vol. 15, no. 2, pp. 169–184, 2011.

[21] G. J. Chowdary, P. Yogarajah, and P. Chaurasia, "MMC-Net: Multi-modal network for cardiac MRI segmentation of ventricular structures, and myocardium," *Institute of Electrical and Electronics Engineers (IEEE)*, 2022.

[22] Y. Chen, L. Wang, B. Ding, Y. Huang, T. Wen, and J. Huang, "Radiologically based automated segmentation of cardiac MRI using an improved U-Net neural algorithm," *Journal of Radiation Research and Applied Sciences*, vol. 16, no. 4, p. 100704, 2023.

[23] A. Ghaznavi, R. Rychtáriková, M. Saberioon, and D. Štys, "Cell segmentation from telecentric bright-field transmitted light microscopy images using a Residual Attention U-Net: A case study on HeLa line," *Computers in Biology and Medicine*, vol. 147, p. 105805, 2022.

[24] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *arXiv preprint arXiv:2211.14830*, 2022.

[25] "PubMed - National Center for Biotechnology Information," https://pubmed.ncbi.nlm.nih.gov/, accessed: 01 10, 2024.

[26] "Heart imaging tests," https://www.nhlbi.nih.gov/health/heart-tests, 2022, accessed: January 14, 2024.

[27] "Cardiac MRI scan," https://www.bhf.org.uk/informationsupport/tests/mri-scan, accessed: January 14, 2024.

[28] C. B. Marcu, A. M. Beek, and A. C. Van Rossum, "Clinical applications of cardiovascular magnetic resonance imaging," *CMAJ*, vol. 175, pp. 911–917, 2006.

[29] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, pp. 155–195, 2016.

[30] M. A. Fatih *et al.*, "Image Analysis," https://bio-protocol.org/exchange/minidetail?type=30&id=9952957, 2021, accessed: January 14, 2024.

[31] A. Andreopoulos and J. K. Tsotsos, "Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI," *Medical image analysis*, vol. 12, no. 3, pp. 335–357, 2008. [Online]. Available: http://jtl.lassonde.yorku.ca/software/datasets/

[32] X. Yang, L. Gobeawan, S. Y. Yeo, W. T. Tang, Z. Wu, and Y. Su, "Automatic segmentation of left ventricular myocardium by deep convolutional and de-convolutional neural networks," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 81–84.

[33] S. Molaei, M. E. Shiri, K. Horan, D. Kahrobaei, B. Nallamothu, and K. Najarian, "Deep convolutional neural networks for left ventricle segmentation," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 668–671.

[34] M. Nasr-Esfahani, M. Mohrekesh, M. Akbari, S. R. Soroushmehr, E. Nasr-Esfahani, N. Karimi, S. Samavi, and K. Najarian, "Left ventricle segmentation in cardiac MR images using fully convolutional network," in *2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 1275–1278.

[35] Z. Gan, W. Sun, K. Liao, and X. Yang, "Probabilistic Modeling for Image Registration Using Radial Basis Functions: Application to Cardiac Motion Estimation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[36] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI," *The MIDAS Journal*, 2009. [Online]. Available: http://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/

[37] C. Constantinides, Y. Chenoune, N. Kachenoura, E. Roullot, E. Mousseaux, A. Herment, and F. Frouin, "Semi-automated cardiac segmentation on cine magnetic resonance images using GVF-Snake deformable models," *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 77, 2009.

[38] Y. Skandarani, N. Painchaud, P.-M. Jodoin, and A. Lalande, "On the effectiveness of GAN generated cardiac MRIs for segmentation," *arXiv preprint arXiv:2005.09026*, 2020.

[39] C. Li, M. Chen, J. Zhang, and H. Liu, "Cardiac MRI segmentation with focal loss constrained deep residual networks," *Physics in Medicine & Biology*, vol. 66, no. 13, p. 135012, 2021.

[40] E. Zhu, H. Zhao, and X. Hu, "Semi-supervised cardiac MRI image of the left ventricle segmentation algorithm based on contrastive learning," *Optoelectronics Letters*, vol. 18, no. 9, pp. 547–552, 2022.

[41] S. A. *et al.*, "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images," *Medical Image Analysis*, 2014. [Online]. Available: http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/

[42] F. Guo, M. Ng, I. Roifman, and G. Wright, "Cardiac Magnetic Resonance Left Ventricle Segmentation and Function Evaluation Using a Trained Deep-Learning Model," *Applied Sciences*, vol. 12, no. 5, p. 2627, 2022.

[43] F. Guo, M. Ng, G. Kuling, and G. Wright, "Cardiac MRI segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors," *Medical Image Analysis*, vol. 81, p. 102532, 2022.

[44] C. Petitjean *et al.*, "Right ventricle segmentation from cardiac MRI: A collation study," *Medical Image Analysis*, pp. 187–202, 2015. [Online]. Available: https://rvsc.projets.litislab.fr/

[45] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2137–2148, 2018.

[46] G. Borodin and O. Senyukova, "Right ventricle segmentation in cardiac MR images using U-Net with partly dilated convolution," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part II 27*. Springer, 2018, pp. 179–185.

[47] Y. Dang, D. Anand, and A. Sethi, "Pixel-wise Segmentation of right ventricle of heart," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 1797–1802.

[48] X. Du, X. Xu, H. Liu, and S. Li, "TSU-net: Two-stage multi-scale cascade and multi-field fusion U-net for right ventricular segmentation," *Computerized Medical Imaging and Graphics*, vol. 93, p. 101971, 2021.

[49] T. Y. Tsai, L. Lin, S. Hu, M.-C. Chang, H. Zhu, and X. Wang, "UU-Mamba: Uncertainty-aware U-Mamba for Cardiac Image Segmentation," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024, pp. 267–273.

[50] R. Xu and I. Oksuz, "Segmentation-aware MRI subsampling for efficient cardiac MRI reconstruction with reinforcement learning," *Image and Vision Computing*, vol. 150, p. 105200, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885624003056

[51] H. Cui, L. Jiang, C. Yuwen, Y. Xia, and Y. Zhang, "Deep U-Net architecture with curriculum learning for myocardial pathology segmentation in multi-sequence cardiac magnetic resonance images," *Knowledge-Based Systems*, vol. 249, p. 108942, 2022.

[52] S. Huang, J. Liu, L. C. Lee, S. K. Venkatesh, L. L. S. Teo, C. Au, and W. L. Nowinski, "An image-based comprehensive approach for automatic segmentation of left ventricle from cardiac short axis cine mr images," *Journal of digital imaging*, vol. 24, pp. 598–608, 2011.

[53] H.-Y. Lee, N. C. Codella, M. D. Cham, J. W. Weinsaft, and Y. Wang, "Automatic left ventricle segmentation using iterative thresholding and an active contour model with adaptation on short-axis cardiac MRI," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 905–913, 2009.

[54] N. C. Codella, J. W. Weinsaft, M. D. Cham, M. Janik, M. R. Prince, and Y. Wang, "Left ventricle: automated segmentation by using myocardial effusion threshold reduction and intravoxel computation at MR imaging," *Radiology*, vol. 248, no. 3, pp. 1004–1012, 2008.

[55] H. Hu, H. Liu, Z. Gao, and L. Huang, "Hybrid segmentation of left ventricle in cardiac MRI using gaussian-mixture model and region restricted dynamic programming," *Magnetic resonance imaging*, vol. 31, no. 4, pp. 575–584, 2013.

[56] J. Folkesson, E. Samset, R. Y. Kwong, and C.-F. Westin, "Unifying statistical classification and geodesic active regions for segmentation of cardiac MRI," *IEEE transactions on information technology in biomedicine*, vol. 12, no. 3, pp. 328–334, 2008.

[57] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac MR images," *Medical image analysis*, vol. 19, no. 1, pp. 98–109, 2015.

[58] S. Ordas, L. Boisrobert, M. Huguet, and A. Frangi, "Active shape models with invariant optimal features (IOF-ASM) application to cardiac MRI segmentation," in *Computers in Cardiology, 2003*. IEEE, 2003, pp. 633–636.

[59] S. C. Mitchell, B. P. Lelieveldt, R. J. Van Der Geest, H. G. Bosch, J. Reiver, and M. Sonka, "Multistage hybrid active appearance model matching: segmentation

of left and right ventricles in cardiac MR images," *IEEE Transactions on medical imaging*, vol. 20, no. 5, pp. 415–423, 2001.

[60] H. Zhang, A. Wahle, R. K. Johnson, T. D. Scholz, and M. Sonka, "4-D cardiac MR image analysis: left and right ventricular morphology and function," *IEEE transactions on medical imaging*, vol. 29, no. 2, pp. 350–364, 2009.

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

[62] F. Kong and S. C. Shadden, "A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net," 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*. Springer, 2021, pp. 287–296.

[63] P. Whig, P. Sharma, R. R. Nadikattu, A. B. Bhatia, and Y. J. Alkali, "GAN for Augmenting Cardiac MRI Segmentation," in *GANs for Data Augmentation in Healthcare*. Springer, 2023, pp. 207–222.

[64] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[65] N. Savioli, M. S. Vieira, P. Lamata, and G. Montana, "A generative adversarial model for right ventricle segmentation," *arXiv preprint arXiv:1810.03969*, 2018.

[66] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[67] H. Zhang and Z. Cai, "ConvNextUNet: A small-region attentioned model for cardiac MRI segmentation," *Computers in Biology and Medicine*, vol. 177, p. 108592, 2024.

[68] H. Aghapanah, R. Rasti, F. Tabesh, H. Pouraliakbar, H. Sanei, and S. Kermani, "MECardNet: A novel multi-scale convolutional ensemble model with adaptive deep supervision for precise cardiac MRI segmentation," *Biomedical Signal Processing and Control*, vol. 100, p. 106919, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809424009777

[69] N. Das and S. Das, "Attention-UNet architectures with pretrained backbones for multi-class cardiac MR image segmentation," *Current Problems in Cardiology*, vol. 49, no. 1, p. 102129, 2024.

[70] N. Subaramani and E. Sasikala, "An attention-based dense network model for cardiac image segmentation using learning approaches," *Soft Computing*, vol. 28, no. 1, pp. 765–775, 2024.

[71] D. Li, Y. Peng, Y. Guo, and J. Sun, "MFAUNet: Multiscale feature attentive U-Net for cardiac MRI structural segmentation," *IET Image Processing*, vol. 16, no. 4, pp. 1227–1242, 2022.

[72] Q. Zheng, H. Delingette, and N. Ayache, "Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow," *Medical image analysis*, vol. 56, pp. 80–95, 2019.

[73] I. B. Ayed, S. Li, and I. Ross, "Embedding overlap priors in variational left ventricle tracking," *IEEE Transactions on Medical Imaging*, vol. 28, no. 12, pp. 1902–1913, 2009.

[74] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. 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*. Springer, 2018, pp. 120–129.

[75] C. Li, Q. Tong, X. Liao, W. Si, S. Chen, Q. Wang, and Z. Yuan, "APCP-NET: Aggregated parallel Cross-Scale pyramid network for CMR segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 784–788.

[76] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[77] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *CoRR*, vol. abs/2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[78] C. B. Wijesinghe, D. Meedeniya, and P. Yogarajah, "Cardiac MRI Segmentation of Ventricular Structures and Myocardium Using U-Net Variants," in *2025 5th International Conference on Advanced Research in Computing (ICARC)*. Belihuloya, Sri Lanka: IEEE, 2025, pp. 1–6.

APPENDIX A

PERFORMANCE EVALUATION OF DL-BASED STUDIES

Study	Dataset(s)	Technique	Pre-processing	Performance
Kamal *et al.* [14]	ACDC 2017, ASC 2018, M&Ms 2020 and LAScarQS 2022	Attention-guided Residual W-Net (ARW-Net)	Image resizing, instance normalization, intensity normalization and data augmentation	For ACDC, Dice scores in ED Phase for LV, RV and MYO are 0.967, 0.950 and 0.905 respectively. In ES phase, the scores are 0.938, 0.895 and 0.923 respectively.
Chen *et al.* [22]	Dataset, provided by The Huaqiao University Affiliated Strait Hospital	U-Net CSP	ROI detection, data normalization and data augmentation	Mean Dice score, considering both ED-ES phase with weighted Cross Entropy loss, is 0.934 (0.043)
Mugahed *et al.* [10]	EMIDEC MRI dataset	End-to-end AI framework with top performing ResU-Net model	image contrast enhancement strategies using the Contrast Limiting Adaptive Histogram Equalization (CLAHE), contrast stretching, intensity level slicing and data augmentation	Overall MIoU: 0.8423
Sharan *et al.* [15]	ACDC 2017	U-Net with VGG encoder and Feature Pyramid Network	Image cropping, intensity normalization using min-max normalization and data augmentation	Dice scores for LV, RV and MYO are 0.958, 0.934 and 0.914 respectively.

Study	Dataset(s)	Technique	Pre-processing	Performance
Ren *et al.* [9]	ACDC 2017	Multi-Task Learning based U-Net (MTL-UNET)		Dice scores for LV, RV and MYO are 0.8818, 0.7244 and 0.8074 respectively.
Hengfei *et al.* [51]	Public MyoPS 2020 challenge dataset	Deep U-Net architecture with curriculam learning and DFM, CAM and SKM modules	Truncate image to keep only gray scale regions, image normalization and data augmentation	Dice (scar region): 0.686, and Dice (edema + scar regions): 0.705
da Silva *et al.* [4]	ACDC 2017	Cascaded approach: U-Net for ROI extraction, FCN for initial segmentation, and a U-Net model for refinement	Image resizing, outlier removal and normalization	ROI extraction using U-Net 1: 0.9368; Dice scores for initial segmentation of LV, RV and MYO are 0.9283, 0.8306 and 0.8474; and Dice scores for refinement of LV, RV and MYO are 0.9230, 0.8304 and 0.8528.
Yutian *et al.* [5]	ACDC 2017	Combined CNN and RNN network that uses Res U-Net as the initial segmentation network and a hierarchical ConvLSTM based recurrent network as the temporal consistency network	Image resampling and data augmentation	Dice scores in ED Phase for LV, RV and MYO are 0.8967, 0.8146 and 0.7260 respectively. In ES phase, the scores are 0.8133, 0.7080 and 0.7656 respectively.

Study	Dataset(s)	Technique	Pre-processing	Performance
Chowdary *et al.* [21]	MS-CMRSeg-2019 challenge dataset and ACDC 2017	Multi-Modal Cardiac Network (MMC-Net)	Image resizing, intensity normalization and data augmentation	Dice scores in ED Phase for LV, RV and MYO are 0.9710, 0.9700 and 0.9450 respectively. In ES phase, the scores are 0.9540, 0.9560 and 0.9600 respectively.
Hengfei *et al.* [13]	Public Left Ventricle Segmentation Challenge (LVSC) dataset	Attention U-Net Architecture with Input Image Pyramid and Deep Supervised Output Layers (AID) network	Image cropping and intensity normalization	Jaccard Index: 0.75; Sensitivity: 0.87; Specificity:0.92
Kong *et al.* [62]	2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms)	A genaralizable, Attention-gated U-Net model and image augmentation using CycleGAN	Image resampling, intensity normalization and data augmentation	Ensemble model performance (Dice scores) for different vendors (A-D): A: 0.888, B: 0.849, C: 0.904 and D: 0.932
Sun *et al.* [12]	ACDC 2017 and SUN09	SAUNet - Shape Attentive U-Net	Z-score-based normalization and data augmentation	Dice scores for LV, RV and MYO are 0.938, 0.914 and 0.887 respectively.
Wu *et al.* [8]	MICCAI 2009 left ventricular segmentation challenge	Composite model: CNN + U-Net	Pre-filter images, downsample images and ROI binary mask creation	Dice-LV: 0.951; Volumetric Overlap Error (VOE)-LV: 0.053; Hausdorff Distance (HD)-LV: 3.641

Study	Dataset(s)	Technique	Pre-processing	Performance
Shibuya *et al.* [7]	Drosophila cell image dataset and Mouse cell image dataset	Feedback U-Net using Convolutional LSTM	Image cropping and data augmentation	MIoU (Drosophila): 0.715; MIoU (Mouse): 0.595
Oktay *et al.* [11]	CT pancreas segmentation problem	Attention gate (AG) model	Intensity normalizaion and data augmentation	Dice-Pancreas: 0.840±0.087

APPENDIX B

PUBLICATIONS

- **Title:** Cardiac MRI Segmentation of Ventricular Structures and Myocardium Using U-Net Variants [78]
- **Conference:** International Conference on Advanced Research in Computing (ICARC)
- **Year:** 2025
- **Reference:** https://ieeexplore.ieee.org/document/10963166