Application of Machine Learning Methods to the Analysis of X-ray Angiography Images

Haorui He

St Hugh's College

A thesis submitted for the degree of

Doctor of Philosophy

Supervised by

Prof. Vicente Grau & Dr. Abhirup Banerjee

Department of Engineering Science, University of Oxford

Hilary Term, 2025

Dedicated to Science

Declaration

I declare that this thesis is entirely my own work and, except where stated, describes my own research.

Haorui He
St Hugh’s College

Acknowledgements

I would like to acknowledge the use of the facilities and services of the Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, and the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. http://dx.doi.org/10.5281/zenodo.22558

I am deeply grateful to my fiancée, Fangqi. Without your support and companionship over the years, I would have struggled through many difficult moments. You have been my anchor, lifting me during times of doubt and making this journey not only bearable but truly meaningful. When I am exhausted, your smile always brightens my world. I know I can be short-tempered when breakthroughs seem out of reach, but you always encourage me to keep trying. When I felt burnt out from research, you always came up with interesting activities to help relieve my stress. Whenever I achieved a breakthrough, you were always there to share the joy with me. I cannot thank you enough for your patience, understanding, and for always being by my side.

A heartfelt thank you to my supervisors, Prof. Vicente Grau and Dr. Abhirup Banerjee. Your unwavering support and guidance over the past years have laid the foundation for this thesis, from my fourth year as an undergraduate at the University of Oxford. Our first meeting in October 2019 marked the beginning of a journey that continued with weekly discussions, igniting my passion for medical imaging and shaping my research every step of the way. You have always been supportive of my new ideas and encouraged me to explore them with curiosity and confidence. Your insightful feedback, patience, and mentorship have not only shaped my research but also helped me grow as an independent thinker. I have learned invaluable skills from you, including academic writing, presentation, and problem-solving, all of which have taught me to be not only a better researcher but a better person.

I am grateful to everyone in the research groups I met at IBME, including Emmanuel Oladokun, Lei Li, He Zhao, Yiying Wang, Chen Chen, Yuling Sang, Felix Wagner, and Marcel Beetz, for the valuable academic discussions and coffee chats we shared. A special thanks to Marcel, whose guidance, experience, and research skills were instrumental in

preparing my first publication.

I also extend my gratitude to Professor Robin Choudhury and his group for their invaluable contributions. Professor Choudhury provided critical clinical insights that shaped my thesis, and I had the privilege of co-authoring with him. Additionally, the data support from his group at the Oxford John Radcliffe Hospital was fundamental to this research.

Thanks to my parents for their emotional and financial support, which allowed me to complete my research journey without worries and with full dedication. Your encouragement, love, and belief in me have been my greatest source of strength, making this achievement possible.

Abstract

Invasive coronary angiography (ICA) is the gold standard imaging modality for diagnosing Coronary Artery Disease (CAD) during cardiac interventions. Accurate segmentation of coronary vessels in ICA could be beneficial for aiding diagnosis and developing effective treatment plans. However, automated vessel segmentation faces multiple challenges, including data scarcity, motion artifacts, uneven contrast distribution, and insufficient feature extraction for semantic analysis. To address these challenges, I first propose a semi-supervised segmentation framework based on a mean teacher model, employing Nested UNets as the backbone. The framework leverages unlabeled ICA images to extract informative features, while an elastic interaction-based loss function helps preserve structural integrity. This approach is trained and evaluated on a dataset collected from the Oxford John Radcliffe (JR) Hospital and demonstrated superior performance compared to state-of-the-art methods. Next, I present a novel Temporal Vessel Segmentation Network (TVS-Net), which fuses sequential ICA frames using a densely connected three-dimensional (3D) encoder and two-dimensional (2D) decoder structure, explicitly disentangling overlapping vessels and analyzing motion in ICA. The model is trained on an ICA dataset comprising 323 samples obtained from the Renji Hospital of Shanghai Jiao Tong University (SJTU), with additional out-of-distribution evaluation conducted on the dataset collected from the Oxford JR Hospital. The results demonstrate superior generalizability compared to six state-of-the-art methods. Finally, I explore both fusion and cascaded approaches by integrating convolutional neural networks (CNNs) with graph neural networks (GNNs) to incorporate geometric features. The proposed method enhances semantic vessel segmentation on a dataset I specifically constructed, using a novel graph generation algorithm and a node label penalty loss, achieving state-of-the-art performance across major coronary vessel branches in the selected view. All proposed methods demonstrate their effectiveness in addressing key challenges, consistently outperforming established benchmarks across multiple evaluation metrics. These advancements underscore the robustness and superiority of my methods for coronary vessel segmentation in ICA images.

List of publications

The following is a list of patents, journal papers, and conference papers published as a result of the research conducted for this thesis.

Patent

- Application Number: GB2402400.2

  Title: Automated vessel segmentation from image sequences

  Applicant: Oxford University Innovation Limited

  Filing Date: 20 February 2024

  Inventors: H. He, A. Banerjee, R. P. Choudhury, and V. Grau

Journal paper

- H. He, A. Banerjee, R. P. Choudhury, and V. Grau, “Deep learning based coronary vessels segmentation in X-ray angiography using temporal information,” *Medical Image Analysis*, vol. 102, p. 103496, 2025.

Conference paper

- H. He, A. Banerjee, M. Beetz, R. P. Choudhury, and V. Grau, “Semi-supervised coronary vessels segmentation from invasive coronary angiography with connectivity-preserving loss function,” *in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2022.
- H. He, A. Banerjee, R. P. Choudhury, and V. Grau, “Automated coronary vessels segmentation in X-ray angiography using graph attention network,” *in International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 209–219, Springer, 2023.

Table of contents

List of figures

List of tables

List of abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
AUPRC	Area Under the Precision-Recall Curve
BFS	Breadth-First Search
CAD	Coronary Artery Disease
CBAM	Convolutional Block Attention Module
CCTA	Coronary Computed Tomography Angiography
ClDice	Centerline Dice
CNN	Convolutional Neural Network
CT	Computed Tomography
Diag.	Diagonal Branch
DS	Deep Supervision
DSA	Digital Subtraction Angiography
ECG	Electrocardiogram
EMA	Exponential Moving Average
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
FP16	16-bit Floating Point
FP32	32-bit Floating Point
GAN	Generative Adversarial Network
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network

ICA Invasive Coronary Angiography

ILSVRC ImageNet Large Scale Visual Recognition Challenge

JR John Radcliffe

LAD Left Anterior Descending Artery

LAO Left Anterior Oblique

LCA Left Coronary Artery

LCX Left Circumflex Artery

MRI Magnetic Resonance Imaging

NLP Natural Language Processing

OCT Optical Coherence Tomography

OM Obtuse Marginal Artery

OOD Out-of-Distribution

PA Posteroanterior

PET Positron Emission Tomography

QCA Quantitative Coronary Angiography

RAO Right Anterior Oblique

RAT Region Attention Transformer

RCA Right Coronary Artery

ReLU Rectified Linear Units

RI Ramus Intermedius

SD Standard Deviation

SGD Stochastic Gradient Descent

SJTU Shanghai Jiao Tong University

SOTA State-of-the-Art

SSM State Space Model

TP True Positive

TVS-Net Temporal Vessel Segmentation Network

VAE Variational Autoencoder

ViT Vision Transformer

VSS Vision State Space

List of symbols

Uncertainty and Stochasticity

κ	Number of stochastic predictions generated for dropout
$\mathcal{H}$	Entropy
μ	Mean probability map obtained by averaging κ softmax outputs
ρ	Exponential Moving Average (EMA) decay rate
ϑ	Threshold value to filter uncertainty map
ζ	Added noise
P	Probability distribution
u	Uncertainty map derived from predictive entropy

Elastic Models and Active Contours

α	Hyperparameter controlling strength of moving boundary dynamics
β	Regularization factor for the Heaviside function
η_{step}	Regularized Heaviside function
γ	Parametric curve in 3D space
ϕ	Level set function representing evolving boundary
$\vec{\tau}$	Unit tangent vector of the curve γ
$\vec{k}$	Unit vector perpendicular to the image
$\vec{r}$	Vector between points
$\vec{v}$	Velocity of the moving curve
$\vec{w}$	Vector field in $\mathbb{R}^3$ related to the elastic energy

$d\vec{l}$ Infinitesimal curve segment

E Energy of the system

Image Topology and Morphology

$\mathcal{B}$ Number of connected components

Model Architecture and Operations

θ Network learnable parameters (weights)

$F(\cdot)$ Feature fusion operation

$f(\cdot)$ Output function of the model

i Depth level (number of down-sampling operations)

j Skip-connection level (number of received skip-connections)

$J(\cdot)$ Join operation: concatenation of feature maps

k Number of levels in the pyramidal UNet++ structure

$R(\cdot)$ Up-sampling (resize) operation

$S(\cdot)$ Softmax operation

$U^{i,j}$ Convolution block in UNet++

$V(\cdot)$ Convolution unit

Loss and Optimization

λ Weighting factor for loss components

$\mathcal{L}$ Loss function

ψ Evaluation metric values

ξ Small constant to avoid numerical instability

$Iter$ Current iteration step

Dataset and Input Parameters

(x, y, z) Spatial coordinates of the image domain

C Number of classes in images

D Dataset

G Gold Standard label

H Height of images

I Input images or data

M Total number of unlabeled data samples

N Total number of labeled data samples

T Total number of temporal frames per sample sequence

W Width of images

Graph and Node-level Learning

χ Grid downsampling exponent used to define the sub-grid size in node construction

δ Connectivity map computed from convolution with skeleton

$\hat{a}$ Predicted class label for nodes in graphs

$\mathcal{A}$ Set of node-level class labels for a graphs

$\mathcal{E}$ Set of edges (connectivity) in the graph

$\mathcal{Q}$ Graph representation

$\mathcal{V}$ Set of nodes in the graph

$\mathcal{Z}$ Adjacency matrix

$\sigma(\cdot)$ Skeletonization function for binary images

υ Geodesic distance threshold used for connecting nodes during edge construction

a Ground truth class label assigned to nodes in graphs

$b^{(*)}$ Node mask

d Number of features per node in the graph

l Interval of iterations after which the graph $\mathcal{Q}$ is regenerated during training

$Mask_B$ Binary mask indicating detected bifurcation points

$Mask_E$ Binary mask indicating detected vessel endpoints

$Mask_M$ Binary mask indicating midpoints on skeleton branches

r, s Indices of nodes in the graph $\mathcal{Q}$

Chapter 1

Introduction

Chapter contents

1.1 Background and Motivation

The human heart plays a central role in the cardiovascular system, responsible for circulating oxygenated blood to tissues and organs and for facilitating the elimination of metabolic waste products. This function guarantees that each cell within the body receives the essential nutrients and oxygen required for survival. Any impairment in the heart or its related vascular system may result in life-threatening conditions, making cardiovascular health one of the most vital concerns in contemporary medicine.

Among many cardiovascular diseases, Coronary Artery Disease (CAD) stands out

as one of the leading causes of death throughout the world, with a growing number of confirmed cases every year [1]. CAD occurs when the coronary arteries, responsible for delivering blood to the myocardium, experience narrowing or obstruction due to plaque accumulation, resulting in stenosis [2]. This limitation in blood flow can cause chest pain, heart attacks, and other significant complications. The precise diagnosis of CAD is imperative for effective intervention and management, highlighting the need for dependable diagnostic methods.

The gold standard for diagnosing CAD and evaluating coronary stenosis is invasive coronary angiography (ICA) during coronary interventions [3]. By delivering high-resolution images of the coronary arteries, ICA allows clinicians to identify blockages and precisely assess the severity of the disease. Automated analysis of ICA, particularly coronary vessels segmentation in ICA, plays a vital role in enhancing this process and forms the foundation for advanced techniques such as quantitative coronary angiography (QCA). QCA enables the extraction of key parameters such as stenosis diameter, lesion length, minimal luminal diameter, and reference vessel diameter, among others [4]. In addition, multiple ICA projections taken from different angles can be used for three-dimensional (3D) reconstruction, providing a more comprehensive visualization of the coronary vascular structure [5]. As the prevalence of cardiovascular diseases continues to rise, rising from 48% in 2014 and projected to be 61% by 2050 in the U.S. alone, there is an increasing demand for automated vessel segmentation methods to provide consistent analysis of ICA images [6, 7].

However, ICA presents inherent challenges such as uneven contrast distribution, motion artifacts, and overlapping structures. With the rapid advancement of deep learning in the medical imaging domain, automated analysis has significantly improved, enabling more accurate and efficient interpretation of complex anatomical features [8, 9, 10]. These techniques leverage the power of machine learning and deep learning to address key challenges relevant to ICA images. To tackle these specific issues, I introduce a combination of innovative approaches, including semi-supervised learning, the integration of temporal information, and the use of graph neural networks (GNNs). The following sections outline the problem statement, research objectives, and contributions of this work, laying the groundwork for a comprehensive exploration of the proposed methodologies and their

results.

1.2 Problem Statement

Although ICA continues to serve as a crucial diagnostic tool for CAD, the segmentation and interpretation of ICA images still present considerable challenges, even for highly skilled clinicians, due to the inherent complexity of X-ray-based imaging. ICA images frequently exhibit cluttered backgrounds with overlapping structures, such as bones and soft tissue, which obscure the regions of interest and diminish vessel contrast [11]. Vessel contrast refers to the difference in pixel intensity between the vessels and the surrounding tissues. This effect makes the detection of small lesions and detailed analysis particularly challenging. Additionally, dynamic variations caused by cardiac and respiratory motions introduce further complications, as the constantly shifting objects disrupt the consistency of vessel segmentation [12]. For example, in Fig. 1.1, the right side of each frame captures the patient's spine, making the vessels in that region challenging to distinguish. Additionally, motion causes the vessels to shift between frames. The vessel indicated by the green arrow is most visible in the second frame, while the vessel marked by the red arrow is most visible in the fourth. None of the frames capture both vessels clearly simultaneously, as the overlapping regions constantly shift due to motion.

Thus, manually segmenting the entire vascular tree without automated methods is labor-intensive and time-consuming, typically requiring about 1 to 2 hours per case [13]. This is especially true for thin vessels, which are challenging to delineate accurately due to their small size, low contrast, and limited visibility. Human error and time constraints further compound these difficulties, often leading to coarse-grained annotations that reduce the quality of available datasets. The scarcity of large, high-quality, open-access datasets for ICA segmentation exacerbates the issue, restricting the development and training of robust automated solutions. Conventional deep learning methods, such as supervised learning, rely heavily on large annotated datasets to achieve high accuracy, and the limited availability of such datasets in ICA severely hampers their applicability and scalability.

Moreover, many deep learning models lack mechanisms to preserve vascular connectivity, frequently producing segmentation errors such as disconnected branches, which

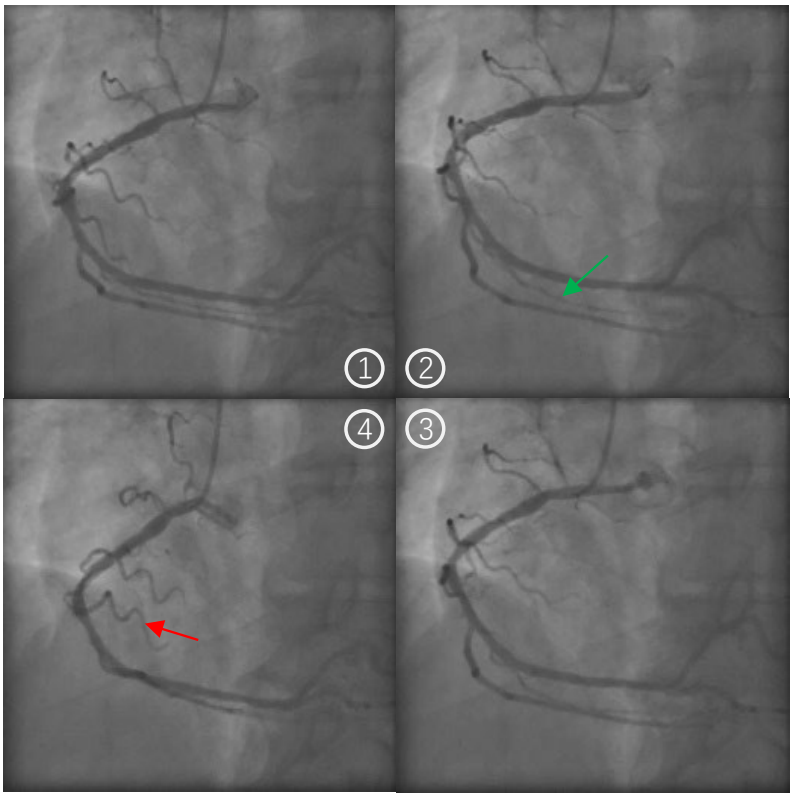


Figure 1.1: A sequence of four consecutive ICA frames focusing on the right coronary artery with the spine appearing in the overlapping region (the right part of every image). The frame order is indicated. Red and green arrows highlight vessels that are maximally visible only in specific frames, where structural overlap is minimal. For example, the green-pointed vessel is not clearly visible in the 1st, 2nd, and 4th frames.

compromise the structural integrity of the vascular tree [14]. Ensuring continuity in the vascular network while avoiding over- or under-segmentation is critical to minimize the risk of misclassifying stenosis and ensure accurate analysis. This limitation is further amplified by the distinct morphology of coronary vessels, where significant differences in vessel width between the main branches and distal branches, combined with motion artifacts, lead to imprecise boundary delineation and reduced generalizability. Even with access to large labeled datasets, these variations introduce complexities that most models struggle to address, especially in regions with overlapping vessels and organs.

Such deficiencies are particularly problematic for downstream tasks like 3D reconstruction of the vascular structure, where preserving the continuity of the vascular network is crucial for accurate assessment [15]. Furthermore, advanced ICA analyses require semantic labeling of vessels to provide a more detailed and interpretable view of coronary artery anatomy and pathology. However, the development of deep learning methods tai-

lored for this purpose remains scarce, leaving a critical gap in the field. Addressing these shortcomings can significantly advance the robustness and applicability of automated ICA analysis, paving the way for improved diagnostic and therapeutic workflows.

1.3 Research Objectives and Contributions

The main objective of this research is to advance the analysis of ICA by developing and applying novel deep learning techniques. Specifically, this study aims to address several key challenges in ICA analysis, including binary and semantic segmentation while maximizing vascular structural integrity. Binary segmentation involves separating the entire vascular tree from the complex background to produce a binary mask, whereas semantic segmentation further partitions the vascular tree into its individual branches.

The first objective of this research is to develop a method that leverages both labeled and unlabeled data for ICA binary segmentation. The availability of large public datasets for ICA is limited due to the complexities involved in their creation, including pathology-specific variability in images and diverse angiographic views. In contrast, unlabeled ICA data is comparatively easier to obtain. This discrepancy demands methods capable of extracting valuable information from a small number of labeled ICAs while effectively utilizing the relatively larger pool of unlabeled ICAs. To address this challenge, I propose a semi-supervised framework incorporating U-Net++ [16] with dropout as the backbone. The framework includes a specialized unsupervised loss that ensures the model produces consistent outputs for unlabeled ICA images, thereby guiding the supervised learning process. Additionally, a tailored supervised loss is employed to reduce disconnections in segmented vessels. This approach minimizes both over-segmentation and under-segmentation while ensuring structural continuity.

After the first objective was completed, ample labeled data became accessible. Thus, the second objective focuses on fully supervised segmentation methods to achieve highly accurate ICA binary segmentation, serving as a benchmark in scenarios where sufficient labeled data is available. Supervised methodologies typically surpass semi-supervised approaches due to their capacity to integrate robust priors into the learning process. This capability is particularly advantageous for tackling challenges such as overlapped

backgrounds, uneven contrast distribution, and motion artifacts that are characteristic of X-ray imaging. To replicate the behavior of experienced ICA annotators, who frequently rely on sequential frames to clarify ambiguities in the target frame, I devise a framework that takes advantage of information from multiple continuous frames to facilitate binary segmentation of the intended frame. This strategy is further enhanced by implementing the supervised loss introduced in the semi-supervised framework, which preserves vascular connectivity while extracting significant information from sequential frames.

The third objective centers on the efficient disentanglement of the structural information inherent within the coronary vascular tree, aiming to enhance segmentation and achieve semantic labeling. This domain remains relatively underexplored, primarily due to the challenges associated with acquiring both binary and semantic labels, along with the inconsistencies observed in vascular structures across diverse patients and pathologies. In response to these challenges, coronary vascular structures are simplified into vessel skeletons with varying widths, subsequently reduced to graph representations by identifying pivotal pixels as nodes and interconnecting them with edges. Using these graph representations, fusion and cascaded pipelines that integrate a convolutional neural network (CNN) and a GNN are developed for semantic segmentation of ICA images, each employing tailored loss functions to enhance performance.

By addressing these objectives, this research aims to fill critical gaps in the current literature and provide innovative solutions to longstanding challenges in ICA analysis, ultimately contributing to the advancement of cardiovascular imaging.

1.4 Organization of the Thesis

Chapter 2 begins with an overview of cardiac anatomy and function, followed by an explanation of the ICA procedure. It then reviews existing research and methodologies for automated ICA analysis, covering traditional vessel analysis techniques alongside advancements in machine learning and deep learning. It discusses challenges such as data limitations, vascular connectivity, and the complexities of coronary morphology, as well as emerging approaches like semi-supervised learning and temporal information analysis. The chapter also highlights the potential of GNNs and semantic labeling techniques to

enhance ICA segmentation and interpretation, providing a foundation for addressing current gaps in the field.

Chapter 3 details the datasets utilized in this study, which serve as the foundation for developing and evaluating the proposed segmentation methods. The Oxford John Radcliffe (JR) Hospital dataset consists of high-quality annotations, albeit with a limited sample size. In contrast, the publicly available Renji Hospital of Shanghai Jiao Tong University (SJTU) dataset offers a significantly larger number of annotated samples. These datasets collectively facilitate the investigation of both semi-supervised and supervised learning methodologies while also supporting robust evaluations across diverse data distributions and the establishment of a semantic labeling dataset. This chapter further outlines the annotation protocols and preprocessing procedures implemented to prepare these datasets for subsequent analyses.

Chapter 4 presents a novel semi-supervised framework for ICA segmentation, drawing inspiration from the mean teacher model. This framework integrates labeled and unlabeled data using a combination of supervised and unsupervised loss functions, where the teacher model extracts uncertainty information from unlabeled data to guide the learning process of the student model. The backbone of this framework is a modified Nested UNet (UNet++), enhanced with a connectivity-preserving elastic interaction energy loss designed to mitigate segmentation discontinuities and maintain vascular structural integrity. This chapter illustrates how the framework effectively utilizes both labeled and unlabeled data to achieve superior segmentation quality while preserving the structural continuity of the vascular tree.

Chapter 5 explores supervised segmentation methods tailored for scenarios with sufficient annotated data. This chapter introduces Temporal Vessel Segmentation Networks (TVS-Net and TVS-Net+), innovative architectures that utilize the temporal information inherent in ICA sequences. These networks incorporate a densely connected three-dimensional (3D) encoder for temporal feature extraction and a two-dimensional (2D) decoder for precise spatial segmentation, enabling seamless fusion of spatial and temporal features. Additionally, applying the connectivity-preserving loss function ensures the structural integrity of the segmented vascular trees. Comprehensive evaluations demonstrate state-of-the-art (SOTA) performance of Temporal Vessel Segmentation Network

(TVS-Net), particularly in metrics focused on structural continuity, highlighting its ability to refine segmentation accuracy and maintain vascular connectivity.

Chapter 6 investigates the structural and semantic analysis of ICA images, building upon the binary segmentation outcomes established in earlier chapters. This chapter explores the use of GNNs for the semantic labeling of coronary vessels. These processes are supported by a specialized labeling interface designed to generate semantic datasets. By modeling the vascular structure as a graph, key nodes and edges are identified to capture essential structural information, which is subsequently extracted as node features. The potential of GNN is initially assessed through its integration with CNN in a fusion architecture for binary segmentation. Subsequently, a cascaded pipeline combining CNN and GNN architectures is introduced, enabling efficient semantic segmentation.

Finally, **Chapter 7** concludes the thesis by summarizing the key contributions, discussing their implications for the field of cardiovascular imaging, and identifying potential avenues for future research.

1.5 Significance and Impact

These comprehensive approaches address the longstanding challenges in ICA segmentation, paving the way for robust and clinically applicable automated analysis methods. The proposed semi-supervised framework offers a transformative solution for clinicians with limited access to large public datasets. By leveraging both labeled and unlabeled data, this framework not only facilitates segmentation in data-scarce environments but also serves as a powerful transfer learning tool. The temporal segmentation of ICA achieves strong performance and generalizability, showcasing the effectiveness of mimicking human annotation behavior by leveraging information across sequential frames. The incorporation of semantic labels into segmentation improves interpretability and utility. By isolating single-vessel masks, clinicians can achieve enhanced visualization and more accurate diagnosis of stenosis locations, leading to more efficient treatment planning and improved patient outcomes. Although the proposed methods primarily involve deep learning architectures, the broader term “machine learning” is also included throughout in this thesis, as some features are manually designed or selected, particularly in the construction of node

features for GNN. The term “machine learning” is therefore intended to capture both the general framework and those more specific application of deep learning methods. Overall, the methodologies and frameworks developed in this study not only address critical gaps in the field of ICA analysis but also hold the potential to streamline and improve the accuracy of coronary artery disease diagnosis.

Chapter 2

Background and Literature Review

Chapter contents

2.1 Introduction

The field of automated invasive coronary angiography (ICA) analysis has made remarkable advancements in recent studies, largely due to the emergence of machine learning and deep learning methods. These technologies aim to overcome the significant limitations of manual ICA analysis, including its labor-intensive nature, susceptibility to human error, and variability among clinicians. Automated ICA analysis offers the potential to enhance diagnostic accuracy and efficiency, making it a crucial area of research in addressing the global burden of coronary artery disease.

This review examines the progression of ICA analysis methodologies, tracing the transition from traditional techniques to modern innovations. Early methods, such as image processing-based vessel enhancement and segmentation, provided foundational tools for ICA analysis. While these techniques delivered initial insights, their inability to manage complex vascular structures and dynamic imaging conditions revealed the need for more advanced approaches. Machine learning and deep learning have since emerged as transformative tools, capable of leveraging large datasets to detect intricate patterns in imaging data and perform tasks such as vessel segmentation, classification, and anomaly detection with unprecedented precision.

The review begins with an overview of cardiac anatomy and the physiological function

of coronary circulation, providing essential context for understanding ICA analysis. It then examines the principles and limitations of invasive coronary angiography, setting the stage for a discussion on the evolution of machine learning and deep learning techniques in medical imaging. The focus shifts to their specific applications and challenges in coronary artery analysis, including issues related to data scarcity, imaging variability, and maintaining vascular structural integrity.

Subsequent sections explore cutting-edge strategies addressing these challenges. These include advancements in semi-supervised learning for data-efficient segmentation, methods leveraging temporal information from video-based ICA sequences, and the application of graph neural networks (GNNs) to capture structural and semantic relationships in vascular trees. Finally, the importance of semantic labeling in ICA analysis is emphasized, with a discussion on the techniques, challenges, and future directions in this domain. Together, these topics provide a comprehensive view of the progress and opportunities in automated ICA analysis, while highlighting areas for future innovation and development.

2.2 Cardiac Anatomy and Function

Understanding the cardiac anatomy and the physiological role of coronary circulation is essential for analyzing ICA. This section explores the structure of the coronary arteries and the pathophysiology of CAD.

2.2.1 Structure of the Coronary Arteries

The coronary arteries, originating from the ascending aorta, ensure that the myocardium receives an adequate supply of oxygen and nutrients to maintain its continuous activity. These arteries are divided into two main branches: the left and right coronary arteries, as shown in Fig. 2.1. The left coronary artery (LCA) bifurcates into the left anterior descending artery (LAD) and the left circumflex artery (LCX). The LAD, which supplies the anterior wall of the heart and the interventricular septum, is critically important due to the extensive myocardial territory it perfuses, with occlusions in this artery often resulting in severe outcomes [17]. The LCX supplies the lateral wall of the heart, while the right coronary artery (RCA) perfuses the right ventricle and portions of the left ventricle [18]. Additional

branches include the diagonal branches (Diag.), obtuse marginal arteries (OMs), and ramus intermedius (RI). The Diag. bifurcate from the LAD, the OMs from the LCX, and the RI typically arise from the left main bifurcation angle, contributing to the perfusion of the inferior and posterior walls of the heart.

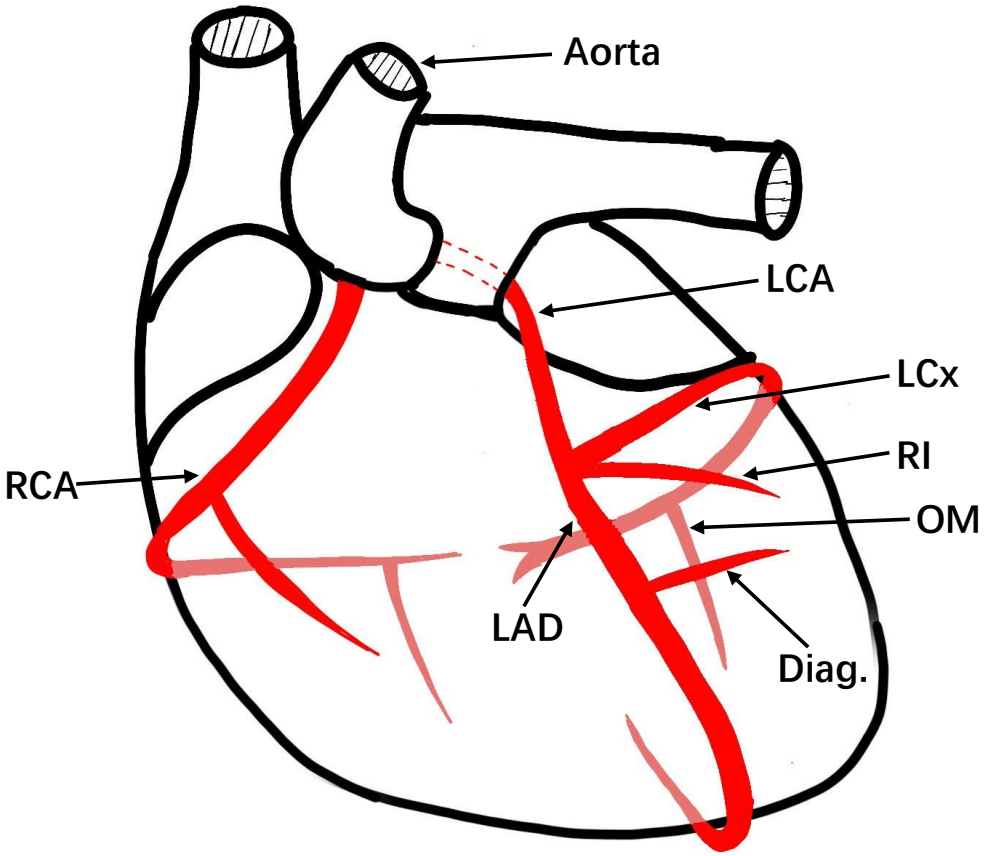


Figure 2.1: Anatomical diagram of the coronary arteries. The figure illustrates the main coronary arteries, including the RCA, LCA, LCx, LAD, Diag., OM, and RI, with arterial paths highlighted in red. The aorta is also indicated as the main vessel supplying blood to the coronary arteries.

However, Fig. 2.1 only shows a general representation of the heart's vasculature. In reality, the specific vasculature can vary greatly from patient to patient, leading to significant semantic differences. For instance, the right inferior (RI) artery may be present or absent in some patients. Additionally, the OM and Diag. branches can show significant variability. The OM branches may be classified as OM1, OM2, and OM3, indicating the presence of three separate obtuse marginal arteries, while the Diag. branches may also appear as Diag.1, Diag.2, and Diag.3, corresponding to multiple diagonal arteries. In

some cases, only one branch of each type may be present, such as OM1 and Diag.1. Furthermore, the length of the RCA can vary, with a longer RCA potentially resulting in a shorter LCx artery, thus influencing the dominance of blood supply to the inferior wall.

The shape of the heart undergoes dynamic changes during the cardiac cycle, with its relaxation (diastole) and contraction (systole) phases facilitating blood circulation throughout the body. Similarly, the morphology of the coronary arteries is also highly dynamic, influenced by both the changing geometry of the heart and the variations in blood flow velocity. During diastole, blood flow can reach velocities of up to 43 cm/s, while during systole, this velocity may decrease to approximately 18 cm/s [19]. In clinical practice, assessing coronary flow velocities and detecting abnormalities in blood flow dynamics are critical for evaluating the severity of coronary artery disease (CAD).

2.2.2 Pathophysiology of Coronary Artery Disease

Coronary Artery Disease (CAD) is characterized by the progressive accumulation of atheromatous plaques within the coronary arteries, as shown in Fig. 2.2. These plaques can reduce or obstruct blood flow, leading to clinical manifestations ranging from stable angina to acute myocardial infarction [20]. The severity of CAD often correlates with the degree of stenosis and the specific arteries affected, with the LAD being particularly critical due to its extensive myocardial supply [21].

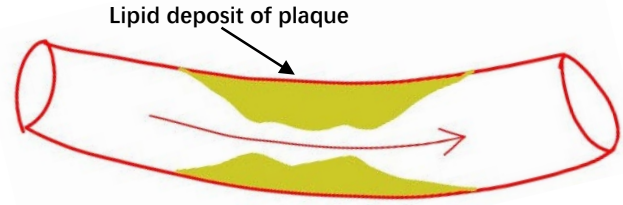


Figure 2.2: Visualization of stenosis: plaque-induced vessel narrowing. The yellow area represents the plaque buildup inside the vessel, which impedes blood flow (indicated by the red arrow).

In addition to atherosclerotic narrowing, coronary artery spasm represents another pathological mechanism of CAD. This condition involves the sudden narrowing of a coronary artery due to the abnormal contraction of vascular smooth muscle, independent of plaque presence [22]. These spasms can lead to temporary ischemia and are often triggered by factors such as cold exposure, stress, or certain medications. In some cases,

coronary spasm may contribute to acute coronary syndromes in the absence of significant stenosis [23].

Risk factors for CAD include both modifiable and non-modifiable components. Modifiable risk factors include hypertension, hyperlipidemia, smoking, diabetes mellitus, obesity, sedentary lifestyle, and dietary habits [24]. Non-modifiable risk factors include age, male sex, and a family history of premature cardiovascular disease [25]. Thus, accurate diagnosis of CAD is essential and relies heavily on imaging modalities like ICA, which provides unparalleled spatial resolution for visualizing arterial lumens and detecting stenoses. However, interpreting ICA images is inherently complex and prone to human error. Automated analysis aims to mitigate these challenges by providing objective, consistent, and precise evaluations of coronary anatomy and pathology [26].

2.3 Invasive Coronary Angiography

ICA is a critical diagnostic tool for visualizing coronary arteries. It is widely regarded as the gold standard for identifying CAD during cardiac interventions. This section discusses the principles of ICA, its advantages, and the limitations and challenges inherent in this imaging technique.

2.3.1 Principles of Invasive Coronary Angiography

ICA is a fluoroscopy-based imaging technique designed to visualize the coronary arteries and identify abnormalities such as stenosis or occlusions. Fluoroscopy is a real-time imaging modality that emits pulsed X-ray photons, delivering ionizing radiation that passes through the patient's body. Different tissues absorb X-rays to varying degrees, and the transmitted X-rays are captured by a detector, which converts them into electrical signals to produce dynamic images [27]. For ICA, the procedure typically begins with the administration of local anesthesia, after which a catheter is inserted into the radial artery and guided to the coronary arteries under fluoroscopic guidance [3]. Once the catheter is correctly positioned, an iodine-based contrast agent is injected to enhance the visualization of the arterial lumen. The high atomic number of iodine increases X-ray attenuation within the blood vessels, thereby improving contrast and enabling high-resolution imaging of the

coronary anatomy [28].

ICA procedures are conducted using a C-arm machine, a specialized X-ray device equipped with a movable arm that rotates around the patient, capturing images from multiple angles, as illustrated in Fig. 2.3. This multi-projection capability, such as posteroanterior (PA) or left or right anterior oblique views (LAO or RAO) and cranial or caudal views, facilitates comprehensive visualization by focusing on different key coronary vessels, including the LAD, LCX, and RCA [29].

The PA view serves as the baseline position, with the X-ray beam directed perpendicular to the patient's chest, just as shown in Fig. 2.3. The RAO view is obtained by positioning the X-ray detector toward the patient's right anterior side, while the LAO view positions the detector toward the left anterior side. Cranial and caudal views involve tilting the detector (sliding the C-arm) toward the patient's head or feet, respectively. For example, when ICA is performed to examine the LCX, imaging angles such as RAO-caudal or RAO-cranial are commonly used. In these views, both the LAD and LCX are typically visible as the main coronary branches.

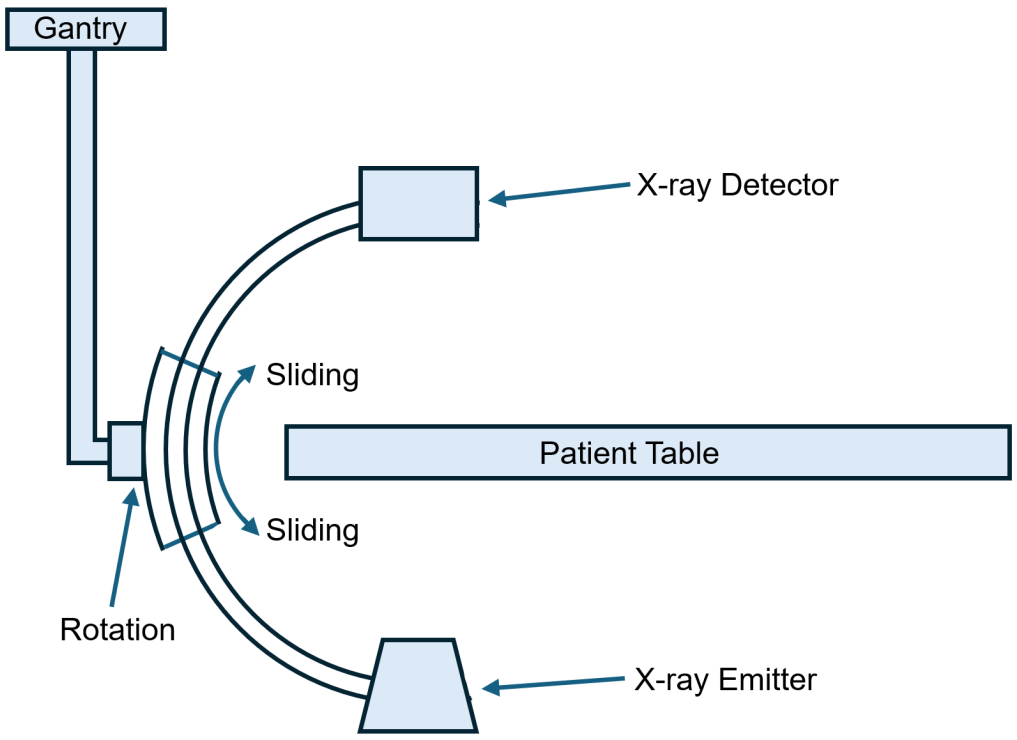


Figure 2.3: A typical ICA system. The patient lies on the table while the C-arm, holding the X-ray emitter and detector, can rotate and slide to capture optimal views of the coronary arteries.

The raw output of ICA consists of video sequences containing approximately 15 to 150 frames, captured at frame rates typically ranging from 10 to 15 frames per second, as illustrated in Fig. 2.4 [30]. This video data is often accompanied by the patient's electrocardiogram (ECG), which provides temporal context through cardiac phases [30]. The imaging detail achieved by ICA is generally high, with the capability to resolve anatomical structures as small as 0.1 to 0.2 mm in diameter [31]. This resolution is primarily limited by the detector's pixel size and patient motion, yet it remains sufficient for precise visualization of the coronary artery lumen and small vascular branches. Additionally, ICA images are affected by several sources of noise. A primary contributor is scatter radiation, which is naturally present in all X-ray imaging procedures [32]. This scattered radiation reaches the detector and degrades image quality by lowering the signal-to-noise ratio, which blurs vessel boundaries and makes accurate segmentation more challenging. The resulting noise follows a Poisson distribution and is particularly noticeable in regions with low photon counts [33].

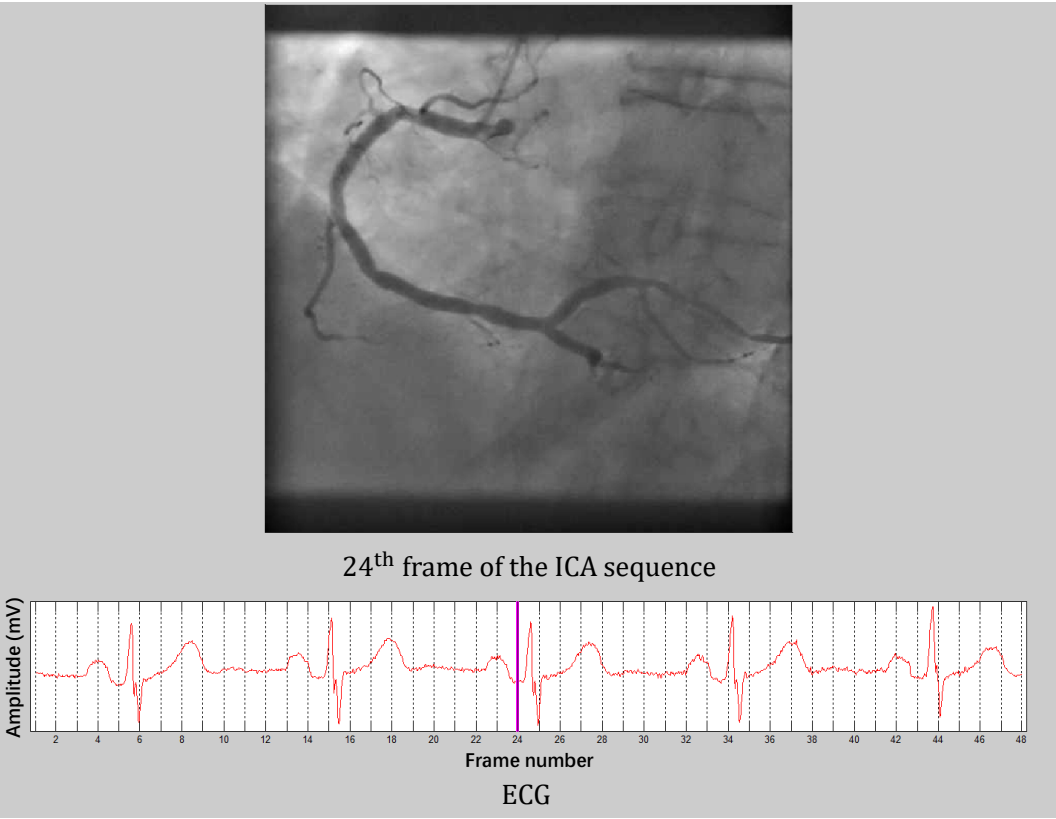


Figure 2.4: A selected frame of the RCA from a 48-frame ICA sequence, extracted at the purple line marker on the corresponding ECG.

The high spatial and temporal resolution of ICA establishes it as the gold standard for assessing coronary artery pathology during interventions. However, despite offering high temporal resolution that captures dynamic changes in vessel motion, this temporal information is rarely utilized in automated analysis, highlighting a promising area for research [34].

2.3.2 Advantages and Limitations of Invasive Coronary Angiography

ICA holds several advantages over other imaging modalities used in CAD, such as coronary computed tomography angiography (CCTA), ultrasound (echocardiography) and cardiac magnetic resonance imaging (MRI). Unlike non-invasive methods, ICA provides direct access to the coronary arteries, making it possible for both diagnostic imaging and therapeutic interventions, such as balloon angioplasty or stent placement, to be performed during the same session. This dual functionality is a significant advantage, allowing for immediate treatment of significant lesions, thereby reducing the need for separate procedures [17].

When compared to CCTA, ICA demonstrates superior accuracy in quantifying the severity of stenosis. This advantage is particularly evident in patients with calcified coronary plaques, where CCTA can produce suboptimal images or overestimate stenosis severity due to beam hardening artifacts [35]. Beam hardening occurs when dense materials, such as calcium, absorb lower-energy X-ray photons more than higher-energy ones. As a result, the CT image is distorted by dark spots or overestimation of the dense areas (blood vessels) [36]. Additionally, ICA is less affected by body habitus, making it suitable for a broader range of patients, including those with high BMI, arrhythmias, or tachycardia, where CCTA may fail to provide adequate diagnostic images [37].

When compared to ultrasound and cardiac MRI, ICA offers superior visualization of the entire coronary vascular tree. Ultrasound, specifically echocardiography, uses high-frequency sound waves for functional imaging of the heart, including assessment of wall motion and valve function [38]. It is a portable, widely available, and cost-effective tool for cardiac evaluation. Cardiac MRI, on the other hand, uses magnetic fields and radiofrequency pulses to assess myocardial perfusion and is particularly well-suited for tissue characterization, such as detecting fibrosis or infarction [39]. However, both ultrasound

and cardiac MRI have significant limitations in imaging coronary arteries. Ultrasound is restricted by its relatively low spatial resolution and dependence on a suitable acoustic window [40]. Cardiac MRI, while powerful for soft tissue analysis, often suffers from motion artifacts due to long acquisition times and challenges in imaging small, fast-moving coronary vessels [41]. Even with the use of specialized contrast agents, these modalities are generally inadequate for detailed visualization of the coronary arteries, where ICA remains the gold standard [40].

Nevertheless, ICA has its drawbacks. Being an invasive procedure, it carries inherent risks associated with arterial access and exposure to ionizing radiation [42]. Additionally, ICA relies heavily on expert interpretation, which introduces subjectivity and may lead to inconsistent diagnostic outcomes, particularly in complex cases [43]. The dynamic behavior of the coronary arteries, compounded by motion artifacts from cardiac and respiratory activity, further complicates image analysis, emphasizing the need for advanced computational techniques to improve accuracy and reproducibility [26].

In summary, while ICA remains the gold standard during interventions for coronary artery imaging due to its high resolution and ability to integrate diagnosis and intervention, advancements in non-invasive methods such as CCTA continue to provide complementary options, with features for each method concluded in Table 2.1. These innovations offer alternative solutions tailored to specific patient needs and clinical scenarios, highlighting the evolving landscape of CAD diagnostics.

Table 2.1: Comparison between different cardiac imaging modalities.

Feature	ICA	CCTA	Ultrasound	Cardiac MRI
Invasiveness	Invasive	Non-invasive	Non-invasive	Non-invasive
Coronary Lumen	Gold Standard	Good	Not visualized	Limited
Vessel Wall	Not visible	Visible	Not visible	Limited
Functional Imaging	Limited	Limited	Strong	Strong
Radiation	Yes	Yes	No	No
Real-time Imaging	Yes	No	Yes	No
Contrast Agent Required	Must	Must	Depends	Depends

2.4 Machine Learning and Deep Learning

Traditional methods for vessel analysis laid the foundation for computational approaches in this field, driving early advancements in vascular imaging. However, these methods faced significant limitations in handling the complexities of intricate vascular structures and imaging artifacts, highlighting the need for more sophisticated solutions. In response, machine learning and, more recently, deep learning have transformed medical imaging by significantly improving accuracy, efficiency, and automation. These methods have outperformed traditional approaches, particularly in coronary artery analysis, by enabling more precise and scalable image interpretation. Machine learning is a broad term for algorithms that learn patterns from data to make predictions without explicit programming [44]. Deep learning, a subfield of machine learning, uses neural networks with multiple processing layers to automatically extract hierarchical features from raw data [45].

This section begins by reviewing traditional vessel analysis methods, focusing on their contributions and limitations. It then explores the evolution of machine learning and deep learning architectures, followed by their applications in coronary artery imaging and the potential challenges of integrating these advanced techniques into clinical practice.

2.4.1 Traditional Statistical and Machine Learning Based Methods

Traditional methods for vessel analysis have been studied for decades and can generally be categorized into filtering-based methods, tracking-based methods, and model-based methods.

Filtering-based methods typically involve enhancing vessel structures in an image by applying filters designed to highlight tubular features [46]. Techniques such as Hessian-based filters have been widely used to enhance vascular structures by analyzing the eigenvalues of the Hessian matrix [47, 48]. These methods effectively detect vessels with consistent intensity profiles but struggle in cases of low contrast or noisy images, which are common in medical imaging [49].

Tracking-based methods focus on tracing vessels by starting from a seed point and following the path of the vessel iteratively. Techniques such as active contour models and region growing are commonly employed in this approach [50]. For example, Bhuiyan et

al. [51] proposed to segment vessels by iteratively expanding regions based on intensity homogeneity. While effective for simpler vessel structures, tracking-based methods often fail in thin overlapping vessels or under significant motion artifacts [52, 53].

Model-based techniques integrate existing knowledge of vascular geometry and structure into their analysis. They utilize geometric models or templates to identify and match vessels within images, with common examples being Level set methods and deformable models. For instance, Wang et al. [54] proposed a level set approach guided by implicit models, which enhances the segmentation process by incorporating vessel-specific geometric constraints. Likewise, convolution surfaces and implicit medial axes can also be used to represent complex vascular topologies [55]. Although model-based approaches are beneficial for maintaining vessel continuity and handling bifurcations, they require careful initialization based on clinical knowledge and are computationally intensive [56, 57].

While these traditional methods laid the groundwork for computational vessel analysis, they face limitations in handling the complex image background and dynamic nature of coronary artery imaging. These challenges, combined with the demand for greater accuracy and automation with reduced human intervention, have driven the adoption of machine learning techniques that utilize large datasets and sophisticated algorithms to overcome the limitations of traditional approaches.

Various machine learning methods have been developed based on different theoretical frameworks, including marginal space learning, random forests, and principal component analysis [58, 59, 60]. In particular, Socher et al. introduced a marginal space learning framework that incrementally searches the parameter space by first estimating vessel position, followed by orientation, and then scale. This hierarchical strategy allows for more accurate vessel detection, particularly in thin vessels, and outperforms traditional filtering-based methods [58]. Gupta et al. developed a multi-stage random forest framework to enhance vessel segmentation in ICA by addressing false positives caused by artifacts such as organ shadows and surgical stitches. Their method applies a cascade of classifiers trained on vesselness and effective margin features to iteratively prune candidate seed points, resulting in the identification of true vessel structures [61]. Similarly, Jin et al. proposed a vessel segmentation method based on local phase congruency

and principal component analysis, where salient vascular features are extracted from the multiscale local phase map [62]. However, machine learning methods depend heavily on handcrafted features, which are typically local and low-dimensional, requiring significant domain expertise for effective feature selection. As a result, these methods struggle to generalize to unseen datasets and are particularly sensitive to noise and variations.

2.4.2 Deep Learning Methods

Over time, deep learning emerged as a transformative advancement, utilizing neural network algorithms where inputs pass through multiple layers of non-linear transformations [63]. These networks optimize weights and biases through backpropagation, a process that computes the gradients of a loss (or objective) function with respect to each parameter, enabling efficient error minimization and resulting in a more automated and scalable learning process [64, 63].

Among all deep learning methods, convolutional neural networks (CNNs), specifically designed for grid-patterned datasets like images, have become the dominant architecture in computer vision at an earlier stage. A CNN consists of three primary components: convolutional layers, pooling layers, and fully connected layers, as shown in 2.5. Convolutional layers extract features by computing the element-wise product between a kernel and corresponding regions of the input tensor. Pooling layers downsample feature maps, reducing spatial dimensions while preserving essential information. Fully connected layers, typically the final layers, combine all nodes with learnable weights to produce the network's output.

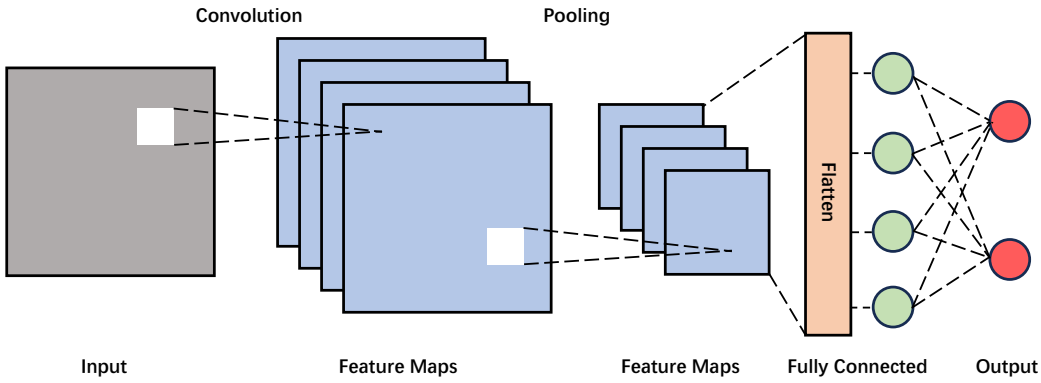


Figure 2.5: Example of a CNN with a compact architecture (one convolution and one pooling).

The first CNN, LeNet, introduced in [65], featured two convolutional layers and was foundational in advancing the field. AlexNet [66], a deeper architecture with 11 layers, achieved remarkable success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [67], showcasing the effectiveness of deeper networks and regularization techniques like dropout.

Subsequent advancements in CNNs concentrated on increasing network depth and refining convolutional block designs to improve performance and functionality. VGG [68] deepened CNN architectures to 16–19 layers, employing smaller convolution kernels (3×3) to achieve state-of-the-art (SOTA) accuracy. GoogLeNet [69] introduced the inception module, which used parallel convolutions of varying kernel sizes within a single layer to capture multi-scale features effectively. ResNet [70] further advanced network depth to 152 layers by introducing residual blocks with skip connections, enabling the network to bypass intermediate layers and mitigate the challenges of training deeper architectures. DenseNet [71] expanded on this idea by connecting each layer to all preceding layers through dense blocks, improving feature reuse and computational efficiency.

These innovations in skip connections and dense architectures significantly enhanced the versatility of CNNs, extending their application beyond image classification to tasks such as image segmentation (e.g., fully convolutional network (FCN) [72] and UNet [16]) and image generation with the introduction of up-sampling layers. However, the increased complexity and computational costs associated with training deeper networks remain an inherent trade-off for achieving higher performance and broader functionality.

To manage the growing complexity of these networks, optimization methods used in backpropagation, such as Stochastic Gradient Descent (SGD) with momentum [73] and Adam optimizer [74], were developed. SGD with momentum accelerates convergence by considering prior gradients, while Adam computes adaptive learning rates using exponentially decaying averages of past gradients and their squares, ensuring efficient parameter updates.

More recently, attention mechanisms have emerged as pivotal innovations in deep learning. Attention layers dynamically assign importance to specific regions in an image, allowing the network to focus on salient features [75]. Originally introduced in natural language processing (NLP) with the Transformer architecture, self-attention enables the

modeling of relationships between all elements in an input sequence, capturing both short- and long-range dependencies [76]. In computer vision, SENet [77] introduced channel-wise attention, adaptively re-weighting feature channels based on their importance, while convolutional block attention module (CBAM) [78] combined spatial and channel attention to enhance feature representations.

The further adaptation of attention to computer vision led to the development of Vision Transformers (ViTs) [79], which treat images as sequences of patches similar to words in NLP. ViT excels in capturing global dependencies, making it particularly effective for segmentation and classification tasks. Building on ViT, Swin Transformer [80] introduced a hierarchical architecture with shifted windows to reduce computational overhead while preserving global dependency modeling, achieving SOTA results on multiple image processing downstream tasks. Nevertheless, all Transformer-based networks necessitate substantial computational resources to attain their objectives, thereby establishing a significant threshold for entry in both research and application.

More recently, hybrid architectures such as Mamba have combined CNNs and Transformers, leveraging state space models (SSMs) to capture extensive contextual information and long-range interactions efficiently [81, 82, 83]. These models integrate CNNs for local feature extraction and Transformers for global feature learning, demonstrating strong performance on complex medical imaging datasets and being more computationally efficient compared to Transformers [82].

2.4.3 Applications of Deep Learning in Medical Imaging

The field of medical imaging has evolved continuously through adaptations and enhancements of various foundational deep learning models such as fully convolutional networks (FCNs), UNet, attention mechanisms, Transformers, and hybrid architectures like Mamba. For instance, FCN proposed by Long et al. [72] employs locally connected layers to perform pixel-wise image classification. A simplified version of its architecture is illustrated in Fig. 2.6. Feng et al. [84] introduced a patch-based FCN with skip connections for retinal blood vessel segmentation, effectively preserving spatial details across layers to improve the accuracy of fine vessel detection. Building on the strengths of FCNs, they also developed a hybrid architecture with short and long skip connections, integrating local

and global contextual information for superior performance in retinal image segmentation [85]. Similarly, Girish et al. [86] tailored an FCN model for optical coherence tomography (OCT) imaging, enabling robust segmentation of intra-retinal cysts by modeling complex tissue structures while maintaining computational efficiency. These advancements set the stage for more unified approaches, such as the work by Maninis et al. [87], who proposed a comprehensive deep learning framework that combines segmentation, lesion detection, and structural analysis to significantly enhance retinal image understanding. However, methods designed for retinal imaging cannot be directly applied to ICA data, due to the greater complexity of backgrounds and significant contrast variability present in ICA images.

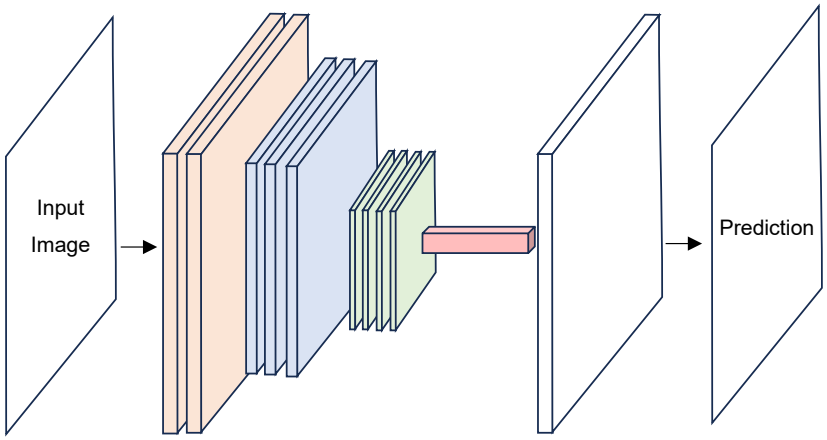


Figure 2.6: Example of a FCN. The orange, blue, and green blocks represent convolutional layers followed by pooling operations. The red block includes a pixel-wise fully connected layer implemented across the channel dimension, followed by a deconvolution layer for upsampling to the desired output size.

Transitioning from FCNs, the UNet architecture, originally engineered for medical image segmentation, has been widely adopted across various imaging modalities due to its effectiveness and flexibility. It consists of an encoder network followed by a decoder network, connected by skip connections that transfer features from the encoder to the decoder [16]. A compact version of this architecture is shown in Fig. 2.7. Recognizing the need for improved accuracy and generalizability, continuous advancements in UNet have led to the development of more sophisticated models. For example, UNet++ enhances segmentation performance by introducing nested and dense skip connections [88], and has demonstrated superior results in ICA segmentation compared to the original UNet

architecture [12]. Building upon these improvements, UNet 3+ further advances the design by integrating full-scale skip connections and deep supervision, enabling more effective multi-scale feature aggregation [89]. Similarly, nnU-Net has emerged as a highly automated framework, optimizing preprocessing, network architecture, and training hyperparameters to adapt seamlessly to diverse datasets [90].

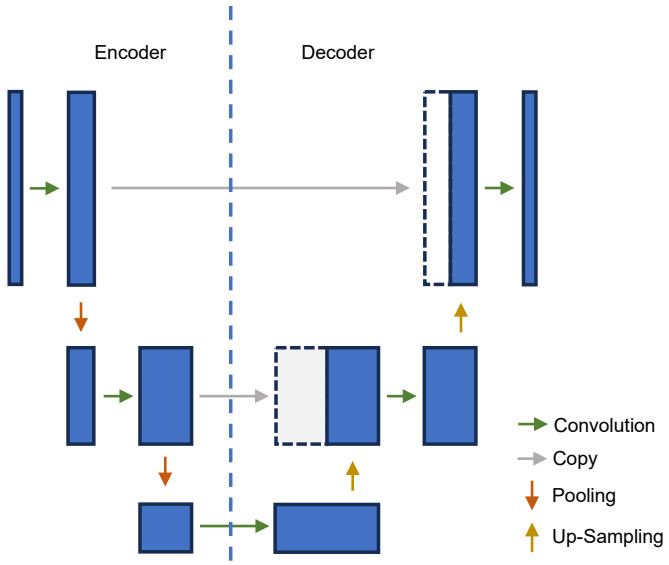


Figure 2.7: Example of a UNet with a compact architecture (3 layers). The colored arrows represent different operations shown in the legend.

As the complexity of medical imaging tasks increased, attention mechanisms and Transformer architectures began to play a pivotal role in advancing the field [91]. For instance, Zhang et al. proposed a dual-branch multi-scale attention network that enables multi-level feature extraction and captures interactive correlations between spatially separated regions, facilitating joint segmentation and quantification of coronary vessels within a unified framework [92]. Additionally, the Region Attention Transformer (RAT) utilizes a region-based multi-head self-attention mechanism to dynamically divide input images into semantic regions, effectively minimizing interference from irrelevant areas and enhancing restoration quality [93]. In medical image segmentation, models like the UNet Transformer integrate a U-shaped architecture with self- and cross-attention mechanisms, allowing for the modeling of long-range contextual relationships and spatial dependencies, which improves vessel segmentation accuracy for complex anatomical structures [94].

Building upon these developments, hybrid architectures like Mamba have recently

demonstrated outstanding performance across many different medical imaging tasks, including image registration, segmentation, and classification [95, 96, 97]. For example, Ruan et al. [96] introduced a purely Mamba-based U-shaped architecture designed for medical image segmentation, utilizing vision state space (VSS) blocks to model long-range dependencies with linear computational efficiency. Its asymmetric encoder-decoder structure, combined with straightforward skip connections, delivers competitive segmentation accuracy while maintaining computational efficiency. However, its application to vessel segmentation remains relatively underexplored.

In parallel, the adoption of pretrained models and transfer learning techniques has become crucial in reducing the computational burden and data requirements for training deep learning models in medical imaging from related tasks, as mentioned in Section 2.4.2. By utilizing models pretrained on large datasets such as ImageNet, researchers have enabled efficient learning in scenarios where labeled medical data is limited [98]. This approach facilitates faster convergence and improved performance in tasks like organ segmentation, lesion detection, and disease classification [99].

Moreover, the demand for real-time processing and deployment on resource-constrained devices has spurred the development of lightweight and efficient models. Techniques such as network pruning and quantization have been applied to architectures like UNet++, reducing the model size and computational requirements without significantly compromising accuracy [88]. These optimizations have made it feasible to deploy deep learning models for medical imaging on edge devices, enabling point-of-care diagnostics and remote monitoring [100].

Through successive advancements, medical imaging has entered an era where traditional and classical machine learning methods have been largely superseded by deep learning approaches, with all the models summarized in Table 2.2. These deep learning models continue to evolve and are increasingly enhanced by emerging technologies, making them more adaptable, efficient, and effective across a wide range of clinical applications.

Table 2.2: Comparison between different approaches for ICA analysis.

Method Type	ICA-Specific Strengths	ICA-Specific Limitations
Filtering-based	Quick enhancement	Poor noise robustness
Tracking-based	Dynamic analysis	Fails in thin vessels
Model-based	Geometry-aware	Initialization dependent
Marginal Space Learning	Accurate in thin vessels	Sensitive to feature selection
Random Forests	Artifact suppression	Handcrafted features
Principal Component Analysis	Effective on local features	Low robustness to variability
CNN-based	Automatic feature extraction	Requires large data
Attention-based	Handles complex backgrounds	Computational complexity
Hybrid (e.g., Mamba)	Efficient + expressive	Not mature for clinical ICA

2.4.4 Challenges in Deploying Deep Learning Models for ICA Analysis

Despite their success in research settings, deep learning models face several challenges when integrated into clinical practice for invasive coronary angiography (ICA) analysis. These challenges primarily include data scarcity, poor generalizability, and issues related to vascular structure disconnection.

One of the most significant limitations is the scarcity of annotated datasets. Training deep learning models for coronary artery analysis requires high-quality annotations provided by domain experts, a time-intensive and resource-demanding process. This scarcity often leads to overfitting, where models learn patterns specific to the training data rather than general trends, leading to poor performance on unseen data. Overfitting is particularly common in medical imaging due to the relatively small dataset sizes, lack of diversity, and confidentiality constraints. To address this, regularization techniques such as dropout are employed during training. By randomly deactivating network nodes, dropout forces the model to learn more robust and generalized representations [101]. Additionally, data augmentation strategies such as image translation, rotation, flipping, cropping, blurring, and the addition of noise help to artificially increase the variability of the training data, improving model resilience to overfitting [102]. More advanced solutions, such as semi-supervised learning and transfer learning, have also been adopted. These techniques utilize unlabeled data or leverage pre-trained models to enhance performance in settings where labeled data is scarce. Generative models, such as generative adversarial network (GAN) and variational autoencoder (VAE), have been utilized to mitigate this issue by generating realistic synthetic datasets [103, 104, 105]. These augmented

datasets improve model robustness in segmentation and classification tasks, particularly for underrepresented populations or rare pathological conditions. Additionally, GANs have been employed to enhance image quality, reduce noise, and simulate pathological variations for training purposes [106]. However, challenges remain in ensuring dataset diversity and achieving a broad representation of clinical scenarios.

Another critical concern is generalizability. Models trained on specific datasets, imaging modalities, or scanner types often fail to perform adequately in new clinical environments due to domain shifts caused by differences in imaging protocols or equipment. This limitation undermines the reliability of even advanced architectures such as CNNs and Transformers when applied to real-world ICA analysis. Evaluating and enhancing the generalizability of these models across diverse datasets (if available) is essential for their successful clinical deployment.

The disconnection of vascular structures in segmentation outputs poses a further challenge for ICA analysis. Addressing this issue is essential for improving segmentation quality and ensuring accurate downstream analyses. Several approaches have been proposed to tackle disconnected vessel issues, primarily in retinal imaging, which can serve as a basis for ICA-specific adaptations. For instance, Lin et al. [107] proposed a GAN-like framework with a discrimination network to guide the segmentation network. Similarly, IterNet was introduced as a model that iterates a mini UNet multiple times with weight sharing and intra-model skip connections, which enhances vessel continuity [108]. However, these implicit methods lack transparent explanations of how connectivity is enhanced, which is crucial for ensuring interpretability and clinical adoption.

Other approaches directly incorporate connectivity-focused loss functions. For example, one method designed a loss function to reduce gaps between predicted and ground-truth vessel boundaries by fusing disconnected segments [109]. Another approach, the centerline Dice (ClDice) loss, was proposed to preserve the skeleton of segmented tubular structures, ensuring topological accuracy while maintaining connectivity of branches [110]. However, while ClDice effectively preserves structural features, its reduced sensitivity to boundary precision may limit its utility in clinically critical regions, such as stenotic areas. Furthermore, connectivity-optimizing loss functions have been introduced to retain structural features like linear continuity and global connectivity while separating background

regions [111, 112]. Although these methods prioritize connectivity, their focus on structural preservation over precise boundary delineation can limit their applicability to coronary vessels, which often feature complex branching and fine-scale details.

These challenges underscore the need for robust strategies to improve data availability, model generalizability, and structural accuracy in vascular segmentation to advance the integration of deep learning models into clinical ICA workflows.

2.5 Semi-Supervised Learning

As emphasized in the previous section, annotated medical imaging datasets, particularly those including ICA segmentation masks, are frequently limited in availability. In contrast, unannotated data is more abundant. This limitation has prompted investigations into the valuable features and information present in large unannotated datasets, a process known as semi-supervised learning.

2.5.1 Fundamentals of Semi-Supervised Learning

Semi-supervised learning offers a middle ground between supervised and unsupervised learning by utilizing large amounts of unlabeled data, which is guided by three core assumptions that relate to the structure of the data distribution [113, 114, 115].

The smoothness assumption states that if two data points are close in the input space, their labels should also be similar [116]. This principle underpins techniques like consistency regularization, where models are trained to produce consistent predictions for perturbed versions of the same data [117].

The cluster assumption suggests that data points within the same cluster in the feature space will likely share the same label [116]. Decision boundaries are therefore encouraged to pass through regions of low data density, ensuring consistent labeling [118].

The manifold assumption posits that high-dimensional data often lie on a lower-dimensional manifold, and points on the same manifold tend to have the same label [116]. This assumption is particularly useful in representation learning, where methods aim to learn embeddings that respect the manifold structure of the data [119].

These three assumptions provide a theoretical foundation for semi-supervised learning methods, allowing them to propagate label information from a small labeled dataset to a larger unlabeled dataset. By exploiting the inherent structure of the data, reliance on manual annotations is reduced while maintaining strong performance in downstream tasks.

2.5.2 Algorithms and Approaches for Semi-Supervised Segmentation

For segmentation tasks in medical imaging, semi-supervised learning has been advanced through various approaches, including consistency regularization, GAN-based methods, and contrastive learning-based techniques. Each approach addresses different aspects of the challenges inherent in using unlabeled data.

Consistency Regularization Methods

Consistency regularization enforces that a model's predictions remain consistent across different perturbations or augmentations of the same input [117]. This principle is particularly effective for segmentation tasks, as it ensures that the spatial structure of the output aligns closely with the input image. Early semi-supervised learning approaches employing consistency regularization focused on enforcing agreement between the outputs of different models or between predictions at different time steps within the same model. These are often referred to as consistency models, including π-model, temporal ensembling, and the Mean Teacher framework [120, 121, 122].

More specifically, the π-model promotes consistency in a model's predictions by minimizing the discrepancy between outputs for original and augmented inputs, thereby ensuring robustness to input perturbations [120]. Temporal ensembling builds on this concept by aggregating predictions across multiple epochs to form a stable target, effectively smoothing noisy pseudo-labels and improving generalization [121]. The Mean Teacher framework extends these principles by employing a student-teacher architecture, where the teacher model generates pseudo-labels for the student to learn from, and the teacher's weights are updated as an exponential moving average (EMA) of the student's weights, ensuring stable and consistent predictions over time [122].

Initially introduced for semi-supervised image classification, the Mean Teacher framework has since been widely adopted for image segmentation tasks, incorporating two

primary perturbation schemes: data transformations like rotation and network perturbations like dropout [123, 124, 125, 126]. In [123], uncertainty estimation is applied by using Monte Carlo dropout [127] to derive segmentation from the teacher model, forcing the student model to align with these consistent outputs. Similarly, in [124], multiple data transformations, including rotation, flipping, and scaling, are employed to encourage transformation consistency between the teacher and student models.

GAN-Based Methods

GANs have gained significant traction in semi-supervised segmentation tasks due to their capacity to generate synthetic data and refine predictions. In GAN-based semi-supervised learning frameworks, a generator produces synthetic segmentation, while a discriminator differentiates between real and synthetic segmentation, as illustrated in Fig. 2.8. For instance, the SegAN framework integrates a segmentation network with a discriminator to refine segmentation outputs, ensuring that the generated segmentations closely resemble ground truth annotations [128]. Moreover, by incorporating complex anatomical constraints within the GAN framework, constrained adversarial training produces anatomically plausible segmentations by emphasizing properties such as convexity and symmetry, which are challenging to encapsulate within a conventional loss function [129]. GANs also contribute to novel data augmentation methods by generating additional training data, thereby enhancing diversity within labeled datasets and improving model robustness [130]. Furthermore, GAN-based methods have been applied in medical imaging to simulate rare anatomical variations, which are commonly encountered in the field. This approach effectively addresses class imbalance in segmentation datasets, where overfitting can easily occur [131].

Contrastive Learning-Based Methods

Contrastive learning emphasizes the acquisition of meaningful representations by contrasting positive and negative pairs of samples, aligning conceptually with the structural principles of the Mean Teacher network [132, 133]. Positive and negative pairs can be derived through data augmentation or region selection from the original data [132, 134]. In segmentation tasks, contrastive learning promotes the alignment of feature embeddings

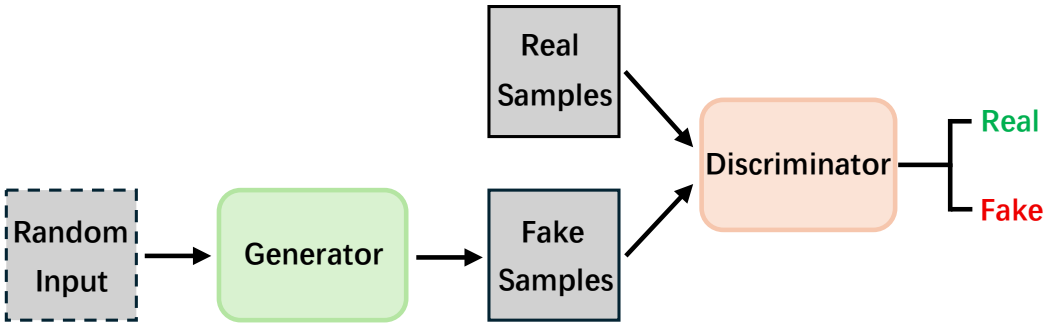


Figure 2.8: Example of a GAN model involves a generator that creates synthetic samples, while a discriminator evaluates each sample to determine whether it is real or artificially generated.

for spatially or temporally adjacent regions while ensuring separation from embeddings associated with unrelated regions. For example, contrastive loss can be employed to enforce the proximity of embeddings from neighboring patches or frames while maintaining distinction from embeddings of non-overlapping regions [134]. Additionally, a local contrastive framework defined over multi-scale feature maps has been introduced, which integrates pseudo-labels and ground truth through a cross-teaching strategy to enhance representation learning [135].

2.5.3 Application in Coronary Artery Analysis

While semi-supervised learning offers general advantages for medical imaging, its importance becomes particularly pronounced in the context of ICA due to various challenges. Annotating ICA frames is not only time-consuming but also prone to inter-observer variability, especially in delineating thin, overlapping coronary vessels and stenosis [14]. Furthermore, ICA suffers from non-uniform contrast enhancement, Poisson-distributed noise, and motion artifacts from cardiac cycles, which degrade image quality and complicate consistent segmentation.

Despite its potential, semi-supervised learning has seen limited application in coronary artery analysis. Among the available approaches, consistency regularization stands out as the most reliable method for achieving segmentation of coronary arteries in ICA, as GAN-based and contrastive learning-based methods present notable limitations [136]. GAN-based methods are particularly effective for data augmentation, generating synthetic ICA images to enhance dataset diversity and address class imbalances. However, their

inherent training instability and the risk of introducing artifacts reduce their suitability for direct segmentation tasks [137]. On the other hand, contrastive learning-based methods improve temporal consistency by aligning features across consecutive frames, which is especially beneficial for video-based angiography. Nonetheless, their dependence on carefully defining positive and negative pairs can hinder their practicality for ICA datasets [138]. Techniques such as the Mean Teacher framework enforce stable predictions across both labeled and unlabeled data, ensuring accurate and robust segmentation even in challenging conditions, such as noisy or low-contrast environments. Incorporating uncertainty estimation methods, such as Monte Carlo dropout, further enhances the reliability of pseudo-labels, making this approach particularly well-suited for ICA segmentation tasks [14].

In conclusion, consistency regularization emerges as an effective approach for coronary artery segmentation in ICA. While GAN-based methods enhance dataset diversity and address class imbalances through data augmentation, and contrastive learning improves temporal coherence in video-based angiography, their limitations make them less suitable for direct segmentation tasks. However, the overall effectiveness of any semi-supervised method remains sensitive to the representativeness of the labeled subset and may vary depending on task complexity and data variability [139].

2.6 Temporal Information Analysis

The dynamic behavior of the coronary arteries is embedded within the temporal resolution of ICA sequences, making temporal information an essential aspect of automated image analysis, particularly for segmentation. Similar to ICA, other imaging modalities, such as ultrasound, MRI, and sequential positron emission tomography (PET), are inherently captured as video sequences. Thus, this section begins by exploring the significance of temporal context in medical image analysis. It then explores techniques for incorporating temporal information into ICA segmentation.

2.6.1 The Role of Temporal Context in Medical Image Analysis

Temporal information was initially employed for motion correction and alignment, serving as a foundational aspect of enhancing medical image analysis. For example, Fessler et al. [140] applied spline smoothing to leverage temporal information, achieving robust tomographic reconstruction and reducing motion-induced inconsistencies to ensure consistency across time frames. Similarly, Atkinson et al. [141] utilized temporal cues to iteratively refine alignment in MR images, improving sharpness and segmentation accuracy even under significant motion distortions.

With advancements in computational technology, methods have evolved to represent data across time, enabling more accurate segmentation by capturing spatio-temporal relationships. For instance, hidden Markov models are used to analyze spatio-temporal brain imaging patterns in serial MRI images [142], leveraging temporal dependencies to track and characterize changes in brain structures over time. In another example, Al-Kadi et al. [143] employed spatio-temporal segmentation techniques in 3D echocardiographic sequences using fractional Brownian motion to model temporal consistency by incorporating statistical properties of motion.

In the field of general computer vision, video segmentation has seen significant advancements, including methods such as region of interest level matching, optical flow-based approaches, and mask propagation techniques [144, 145, 146, 147]. These developments have inspired new methodologies in medical video segmentation as well [148, 149]. However, most of these techniques are typically applicable only to data types where segmentation masks can be obtained for every frame in the sequence, a scenario that is impractical for imaging modalities like ICA due to the time-consuming nature of segmenting even a single frame. Thus, temporal information needs to be specially designed so that it can fuse with spatial information in feature extraction. For instance, in digital subtraction angiography (DSA) for cerebral vessel segmentation, Su et al. [150] introduced a spatio-temporal UNet that simultaneously integrates spatial and temporal features through a Temporal Learning module, enabling cohesive decoding and effective learning of spatial-temporal dynamics. Alternatively, Xie et al. [151] proposed DSANet, which separates spatial and temporal feature extraction using a Temporal Former module

to capture temporal relationships and a Spatio-Temporal Fusion module to merge the extracted features. This approach employs a cascade decoding framework to enhance segmentation. However, compared to ICA images, DSA for cerebral vessels involves simpler background structures and fewer motion artifacts, presenting distinct challenges for adapting such methodologies to ICA segmentation.

While these approaches originate from other medical imaging modalities, their relevance to ICA analysis is significant. Due to the limited availability of annotated ICA video datasets, direct research on temporal segmentation in ICA remains relatively scarce. Therefore, exploring temporal strategies in related imaging contexts provides essential insight into techniques that could be adapted or extended for ICA applications.

2.6.2 Techniques for Leveraging Temporal Information in ICA Segmentation

The use of optical flow for video segmentation has been applied to ICA segmentation due to its ability to capture motion dynamics, as demonstrated by a multi-stage UNet architecture that integrates low-level binary segmentation with optical flow [152]. However, to fuse spatial and temporal features more effectively, specialized models have been developed to address the unique challenges of ICA segmentation. A traditional UNet enhanced with temporal fusion convolution and channel attention mechanisms was proposed to integrate temporal features into spatial representations; however, it did not prioritize preserving the structural integrity of the vascular tree [34]. Other approaches have aimed to enhance the integration of temporal information into segmentation frameworks, tailoring methods to the complexities of dynamic angiographic sequences. A Semi-3D UNet was introduced for coronary vessels segmentation in angiography videos, designed to combine spatial and temporal feature extraction by processing an odd number of consecutive frames centered on the target frame [153]. This approach incorporated temporal context effectively but was limited in its capacity to generalize across varying sequence lengths. Similarly, a tri-pathway FCN for three-frame ICA segmentation was proposed, utilizing an influence matrix to decode temporal information more effectively [154]. All these advancements highlight the importance of developing models that effectively integrate spatial and temporal features to improve segmentation performance. Nevertheless, addressing key challenges

in ICA segmentation, such as maintaining continuity in vascular structures, remains crucial.

2.7 Graph Neural Network

Graph neural networks (GNNs) have emerged as powerful tools for learning representations of data with complex relationships and non-Euclidean structures, making them particularly relevant for medical imaging tasks. By modeling data as graphs, where nodes represent entities (e.g., anatomical regions, vascular segments), and edges encode relationships (e.g., connectivity, adjacency), GNNs offer a natural way to capture the intricate spatial and structural patterns inherent in medical data. This section explores the basics of GNNs and their role in medical imaging. It then reviews current applications of GNNs and discusses their potential for improving ICA segmentation and classification by modeling vascular structures and connectivity.

2.7.1 Basics of Graph Neural Networks and Their Relevance

The core idea of GNNs lies in leveraging message-passing mechanisms to update node representations by aggregating information from their neighbors, enabling the capture of both local and global dependencies [155]. Scarselli et al. [156] first formalized GNNs, introducing a framework that computes node embeddings through iterative message-passing. This foundational work paved the way for subsequent advancements, such as the graph convolutional network (GCN) by Kipf et al. [157], which streamlined graph convolutions to make them scalable and efficient for large datasets, as illustrated in Fig. 2.9. Later, Hamilton et al. [158] expanded on these ideas with GraphSAGE, a method that generates node embeddings for unseen nodes by sampling and aggregating features from their neighborhoods. This innovation allowed GNNs to handle dynamic and evolving graphs effectively. The introduction of attention mechanisms further enhanced GNNs. Velickovic et al. [159] proposed graph attention network (GAT), which uses self-attention to assign different weights to neighbors, focusing on the most relevant connections. This improves performance on heterogeneous and sparse graphs. Collectively, these developments demonstrate the power of GNNs in analyzing structured data.

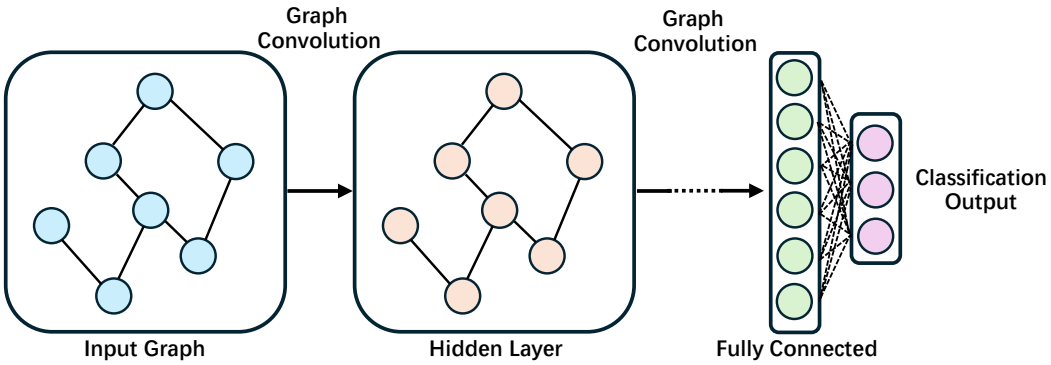


Figure 2.9: Example of a GCN model for graph classification. The structure is similar to a CNN, with the data format replaced by graphs (nodes and edges).

2.7.2 GNNs in Medical Imaging: Current Applications

GNNs for Classification

GNNs have been widely applied to classification tasks, leveraging graph-based representations to extract meaningful features from medical images. For example, in functional brain imaging, Ktena et al. [160] utilized GCNs to learn distance metrics for brain connectivity networks, improving classification accuracy for neurological disorders. Similarly, Parisot et al. [161] modeled patient networks derived from brain imaging to predict autism spectrum disorder, demonstrating the potential of GNNs in disease diagnosis. Another significant contribution comes from the RA-GCN framework, proposed by Ghorbani et al. [162], which addresses the issue of imbalanced datasets in disease prediction. By re-weighting class representations and employing adversarial training, RA-GCN improves classification performance for underrepresented classes, making it particularly effective in medical imaging scenarios where data imbalance is common.

GNNs for Segmentation

GNNs have shown significant effectiveness in medical image segmentation by capturing complex spatial relationships. ViG-UNet, for instance, integrates graph neural networks within a U-shaped architecture, representing images as graphs to model both local and global dependencies [163]. This method has demonstrated superior performance compared to traditional UNet variants on various datasets, achieving better segmentation of intricate structures [163]. In the field of histopathology, a GNN-based network for

tissue semantic segmentation uses graphs constructed from image regions, leveraging spatial and semantic relationships to ensure accurate and consistent segmentation [164]. Furthermore, Zhang et al. [92] introduced a joint fully convolutional and graph convolutional network, which combines pixel-level features from fully convolutional networks with graph-based reasoning. This hybrid framework enhances segmentation in pathology images, particularly with weak annotations, by using graph structures to infer missing details effectively [92].

2.7.3 Potential of GNNs in ICA Analysis

Despite their success in other imaging modalities, Graph Neural Networks (GNNs) are rarely applied to vessel segmentation, particularly for coronary artery segmentation in ICAs. Their use has primarily been explored in retinal vessel segmentation and vessel analysis in 3D computed tomography (CT) images. Although these medical images are typically represented as pixel-wise grid data, the vascular structures can be extracted into graph representations [165]. This transformation allows GNNs to leverage topology and connectivity patterns that are often challenging for CNNs to capture, especially given the high variability in pixel-wise features of vascular structures between patients, as discussed in Section 2.2 [166]. Thus, these prior studies highlight the potential of GNNs for coronary artery analysis, aiming to improve binary and semantic segmentation performance.

For instance, Vessel Graph Networks (VGN) integrate GNNs with CNNs to enhance vessel segmentation by capturing vascular connectivity. This approach has demonstrated success in retinal and coronary artery datasets by combining local and global structural features, improving both segmentation accuracy and vascular tree continuity [167]. Building on this, GNNs have been applied to vessel segmentation using link prediction frameworks. One study constructs a graph where nodes represent vessel segments and link prediction is employed to identify connections, ensuring accurate segmentation and structural consistency [168].

Graph-based approaches have also been used for reconstructing 3D vessel morphology. A GAT-based pruning method was proposed to reconstruct 3D liver vessel morphology from contrasted CT images, refining the segmentation by removing irrelevant structures and preserving critical vessel connectivity [169]. Additionally, a 3D graph-

connectivity constrained network has been introduced for hepatic vessel segmentation, leveraging graph structures to enforce connectivity constraints during segmentation. This approach ensures topological accuracy and maintains the continuity of the vascular tree, critical for applications requiring precise vessel morphology [165].

These studies highlight the potential of GNNs for ICA analysis by addressing challenges such as vascular disconnection and complex branching patterns. Adapting these techniques to coronary artery segmentation could significantly enhance segmentation accuracy and improve structural consistency in ICA analysis, paving the way for more robust and automated approaches in clinical practice.

2.8 Semantic Labeling

Semantic labeling in medical imaging extends beyond segmentation by assigning meaningful labels to anatomical structures or pathological regions, enabling clinical interpretation and decision-making [170]. For coronary artery analysis, semantic labeling involves accurately identifying and labeling specific vessels and branches. This section explores the foundations of semantic labeling in medical imaging, current techniques for coronary artery analysis, and the integration of advanced methods like GNN to enhance labeling accuracy.

2.8.1 Overview of Semantic Segmentation in Medical Imaging

While numerous advanced semantic segmentation models exist in the field of computer vision, most require extensive datasets for effective training. This poses a significant challenge in medical imaging, where annotated data is often limited and domain-specific. In such cases, semi-supervised learning becomes essential. However, when sufficient annotated data is available, such as in carefully curated clinical datasets, fully supervised learning remains a robust and reliable approach. As outlined in Section 2.4, a range of machine learning methods can be applied to semantic segmentation tasks in these cases, particularly those using three primary architectures: UNet-based methods, attention-based methods, and GAN-based methods [171].

UNet has established itself as a cornerstone architecture for medical image segmen-

tation due to its encoder-decoder structure and skip connections, which preserve spatial details. Holste et al. [172] employed a UNet-style architecture for multi-class segmentation of lung organs in chest radiographs, testing various loss functions to mitigate the effects of class imbalance in pixel-wise labels. Similarly, Ahmad et al. [173] developed a multi-scale UNet to address the scale variation problem in MRI-based heterogeneous organ segmentation, incorporating a hierarchical block between the encoder and decoder to enable multi-scale feature extraction.

Attention mechanisms have further enhanced semantic segmentation by enabling models to focus on critical image regions while suppressing irrelevant information. For example, Qin et al. [174] introduced an Autofocus block based on dilated convolutions and attention mechanisms, which adaptively adjusted the size of the effective receptive field. This approach ensured scale invariance by sharing weights across parallel convolutions. In another work, Sinha et al. [175] proposed a self-guided attention mechanism to integrate local features with their corresponding global dependencies, allowing the network to suppress irrelevant regions and focus on discriminative features.

GANs have also shown promise in semantic segmentation tasks. Dai et al. [176] proposed a structure-correcting GAN that employs a critic network to learn structural regularities from human physiology, enabling the model to differentiate between accurate and flawed semantic segmentation maps. Similarly, Rezaei et al. [177] utilized a conditional GAN to train a UNet in conjunction with an adversarial network, enhancing the segmentation accuracy of the UNet by leveraging the adversarial learning paradigm.

Despite their success, these methods are generally designed to segment pathology regions with regular shapes, such as tumors or lesions. In contrast, vessel segmentation in ICA presents unique challenges, including irregular vessel alignment, variable topology, and scale variations. While the discussed architectures provide a solid foundation, adapting them to address the complexities of ICA requires incorporating domain-specific modifications and constraints to improve performance in the semantic labeling of vascular structures.

2.8.2 Techniques for Semantic Labeling of Coronary Arteries

Semantic vessel segmentation methods have rarely been applied to ICA and are primarily developed for retinal images, where they focus on separating arteries and veins from the overall vascular tree. The availability of standardized public datasets in retinal imaging has facilitated consistent benchmarking and the rapid development of innovative techniques. Early work often adapted UNet and GAN-based architectures to address the specific challenges of retinal image segmentation [178]. For instance, Chowdhury et al. [179] proposed a UNet-based architecture with a multiscale extraction encoder and a self-attention decoder, effectively handling the complex morphology and anatomical variability of retinal arteries and veins. Similarly, Chen et al. [180] introduced a specialized GAN that learns topological features using a pretrained topology-ranking discriminator and captures width information through a width-specific loss function by distinguishing width maps of multi-scale dilations.

More recently, CNN-GNN fusion methods have gained prominence by combining the feature extraction capabilities of CNNs with the structural modeling strengths of GNNs. These hybrid approaches have demonstrated significant improvements in classification accuracy and the preservation of vascular topology [178]. For example, Noh et al. [181] proposed a U-shaped GNN that integrates features generated by two parallel CNNs applied to binary segmentation maps, emphasizing vessel connectivity during classification. Mishra et al. [182] designed a voxel-based graph generation algorithm that combines input data and CNN-extracted feature maps to preserve vessel topology while improving classification accuracy. Furthermore, Xu et al. [183] introduced a cascading CNN-GNN architecture, employing a GNN as a connectivity refinement component to enhance the semantic segmentation of retinal arteries and veins. While these methods offer valuable insights, their reliance on voxel neighborhood relationships for graph generation presents challenges. Such approaches often generate either overly dense graphs that are computationally expensive or sparse graphs that fail to adequately represent vascular structures, limiting their applicability to complex vessel networks.

In ICA, where public datasets and standardized benchmarks remain scarce, GNN-based approaches are still emerging. However, graph structures have been applied to

vessel segmentation without explicitly employing GNNs. Zhao et al. [184] developed a framework that models vessel connectivity as a graph, focusing on preserving vascular tree continuity and addressing segmentation challenges without relying on GNN-specific operations. Correspondingly, Zhao et al. [185] used a graph-matching network to explore correlations between different graphs for the generation of vascular semantic labels. These approaches highlight the potential of graph-based methodologies for semantic labeling in ICA and set the stage for integrating GNNs into ICA analysis in the future.

While advancements in retinal vessel segmentation provide a strong foundation, their adaptation to ICA remains complex due to significant differences in imaging modalities. ICA introduces additional challenges, such as higher motion artifacts, more intricate vascular structures, and a greater number of semantic target classes. Addressing these domain-specific challenges will require robust methods tailored to ICA's unique characteristics, paving the way for improved semantic labeling in this domain.

2.9 Conclusion

This chapter starts with fundamental knowledge behind coronary vessels and ICA imaging modality, then provides a comprehensive review of advancements in automated ICA analysis, emphasizing the progression from traditional methods to cutting-edge machine learning and deep learning approaches. While foundational, traditional vessel analysis techniques are limited in their ability to address the complex and dynamic nature of coronary artery imaging. The introduction of machine learning and deep learning methods has significantly improved the accuracy, efficiency, and scalability of ICA analysis. This chapter has focused on the challenges posed by the scarcity of annotated datasets, imaging variability, and maintaining vascular structural continuity. Semi-supervised learning has emerged from all deep learning methods as a promising and straightforward solution for analysis with insufficient data, leveraging both labeled and unlabeled data to improve model generalization. Temporal information, often underutilized in ICA, even with sufficient data, presents another avenue for enhancing segmentation by incorporating spatiotemporal dynamics to disentangle vessels in motion. Similarly, the application of GNNs has demonstrated potential for capturing vascular topology in segmentation tasks. Finally,

semantic labeling, an essential component of ICA analysis, extends beyond segmentation by assigning meaningful labels to anatomical regions. While methods developed for other imaging modalities provide a solid foundation, adapting these techniques to ICA requires addressing its unique challenges, such as motion artifacts, complex vascular structures, and scale variability. The integration of GNNs and hybrid architectures holds significant promise for overcoming these obstacles.

Despite the advancements achieved, significant challenges persist in developing robust, interpretable, and clinically applicable models for ICA analysis. These challenges include segmentation with limited annotated datasets, effective extraction of temporal information from video sequences, and accurate semantic labeling of vascular structures. Moreover, the black-box nature of many deep learning models limits their adoption in clinical settings, where transparency and explainability are crucial for clinical decision-making. Therefore, future research should prioritize domain-specific adaptations of current techniques and focus on improving the generalizability of models across diverse clinical environments with explainable AI techniques. Overcoming these obstacles will enable automated ICA analysis to revolutionize the diagnosis and management of coronary artery disease, ultimately enhancing patient outcomes and advancing cardiovascular imaging.

Chapter 3

Datasets

Chapter contents

3.1 Introduction

The quality and quantity of datasets play a critical role in the development and evaluation of automated medical image analysis, and in particular, segmentation methods. In the context of invasive coronary angiography (ICA), precise segmentation of coronary vessels is essential for quantitative assessments and clinical decision-making. This chapter introduces the two datasets utilized in this study: the JR dataset (D_1) sourced from the Oxford John Radcliffe (JR) Hospital and the SJTU dataset (D_2) derived from publicly available data from the Renji Hospital of Shanghai Jiao Tong University (SJTU). These datasets, containing ICA image sequences annotated for vascular segmentation, provide

a robust foundation for developing and validating the proposed segmentation methods.

The JR dataset (D_1), comprising 120 ICA sequences with 60 manually annotated frames, is smaller but includes high-quality annotations and synchronization with electrocardiogram (ECG) signals. It is used for semi-supervised learning and as an out-of-distribution (OOD) dataset to evaluate models trained on the SJTU dataset (D_2). The SJTU dataset, with 323 annotated frames from 120 ICA sequences, offers a larger sample size suitable for supervised training and in-distribution testing.

These datasets complement each other, enabling the exploration of both semi-supervised and supervised learning approaches while supporting robust evaluation across different data distributions. This chapter details the composition, annotation protocols, and pre-processing steps for both datasets, establishing the foundation for the experiments.

3.2 JR Dataset D_1

The JR dataset employed for this study was procured from the Oxford JR Hospital. It comprises 120 DICOM format sequences of consecutive ICA images collected from 26 patients undergoing coronary examinations related to cases with suspected coronary stenosis, acquired using the Siemens Artis icono angiography system. All patients provided informed consent as part of the Oxford Acute Myocardial Infarction (OxAMI) study protocol, with patient information omitted. These sequences are categorized into 73 sequences captured for the right coronary artery (RCA) and 47 sequences for the left anterior descending artery (LAD). Each sequence contains up to 45 frames, corresponding to multiple phases across several cardiac cycles. The ICA image sequences were synchronized with the patient's ECG during acquisition, enabling precise identification of the end-diastolic phase within each cardiac cycle. The ECG signal is stored as raw digital amplitude values, which can be linearly converted to physical units (e.g., millivolts) using a scaling factor.

Among the multiple end-diastolic frames captured across cardiac cycles, the frame with the most clearly visible vascular structures is selected as the target frame for manual annotation (as shown in Fig. 3.1). The end-diastolic frame corresponds to the point at which the left ventricle attains its maximum volume, typically occurring near the peak of the R-wave in the ECG signal, which is the most prominent peak in the cardiac cycle [186].

This phase is preferred for analysis because the heart is in its most relaxed state, which reduces motion artifacts and allows for optimal visualization of the coronary vessels.

Furthermore, as discussed in Section 2.3, the noise in ICA images follows a Poisson distribution. Frames with poor contrast enhancement, which are often observed at the beginning and end of the sequence, tend to exhibit higher perceptual noise due to a lower contrast-to-noise ratio, even if the photon count is relatively high. By selecting a frame that offers both high clarity and minimal interference from noise or overlapping structures, this approach ensures that the annotated target frame accurately represents the vascular anatomy. This reduces potential errors caused by motion blur, noise, or anatomical ambiguity in other frames.

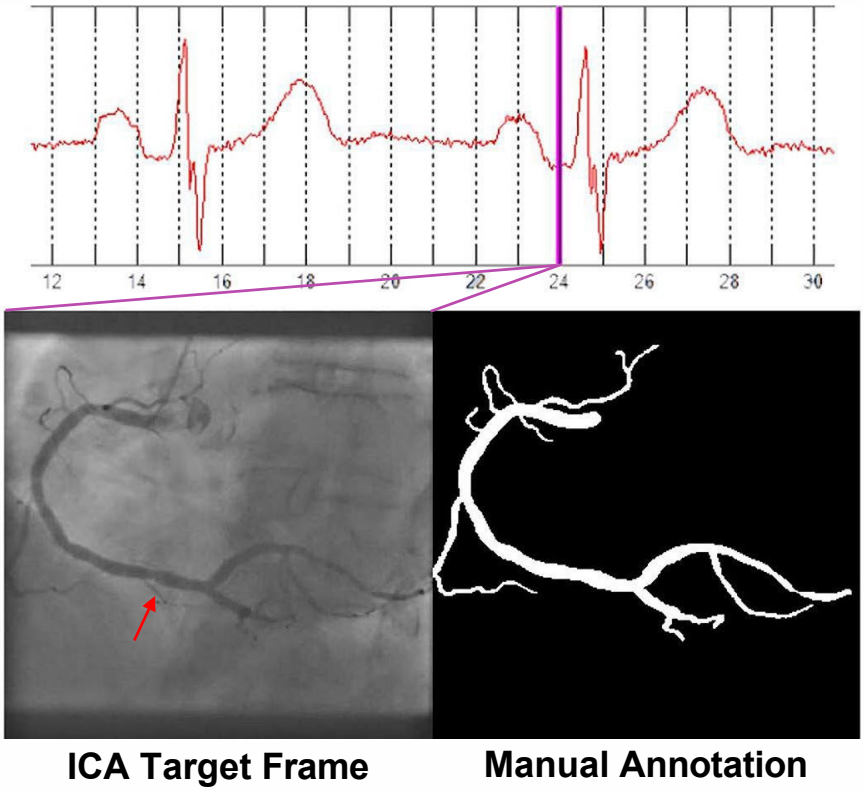


Figure 3.1: Selected ICA frame (the pink lines) from a sequence for RCA with the patient's real-time ECG at the top and its corresponding manual annotation. The red arrow highlights the extra-thin vessel that is excluded in this frame.

The complete dataset comprises 120 ICA frames, with 73 RCA frames and 47 LAD frames. These images have dimensions of $H \times W = 512 \times 512$, where H and W represent the height and width of the images, respectively. The annotation process for this dataset involved three experts. A domain expert selected and manually annotated a subset of 60

frames, including 27 RCA and 33 LAD images, using binary masks to label the vascular regions. These annotations were first reviewed by another clinical expert, followed by a more senior clinical expert who performed additional refinement to ensure the highest level of accuracy and consistency in delineating the coronary vessels. The entire process, guided by a clearly defined annotation protocol, involved manually delineating boundaries for vascular regions, as illustrated in Fig. 3.1 for the RCA and Fig. 3.2 for the LAD. This protocol focused on prioritizing clinically significant vascular structures while intentionally excluding extra-thin vessels, as indicated by the red arrows in Figs. 3.1 and 3.2. The exclusion of these vessels was intended to strike a balance between annotation precision and efficiency. Since extremely thin vessels, typically only 2 to 3 pixels wide, have a limited impact on overall segmentation performance but demand significantly more time to annotate manually, omitting them ensured that the final labels remained both accurate and practical for the study's objectives.

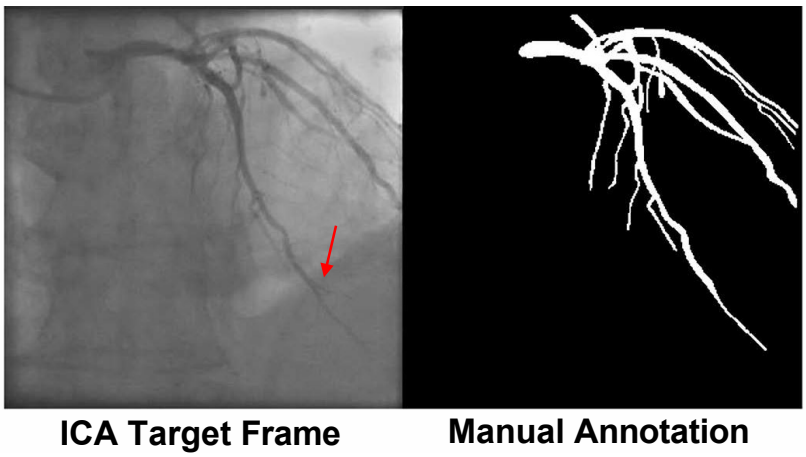


Figure 3.2: Selected ICA frame of the LAD and its corresponding manual annotation. The red arrow highlights the extra-thin vessel that is excluded in this frame.

The annotation process, starting with target frame selection, was supported by a custom tool developed by former students, shown in Fig. 3.3. It allows users to designate vascular areas by drawing red lines along vessel contours. The red contours in Fig. 3.3 extend beyond the boundaries of the zoomed ICA image to provide a broader visualization area.

The tool is designed to maximize boundary precision by offering multiple methods for creating and adjusting contours. Users can manually pin each red point in the contours

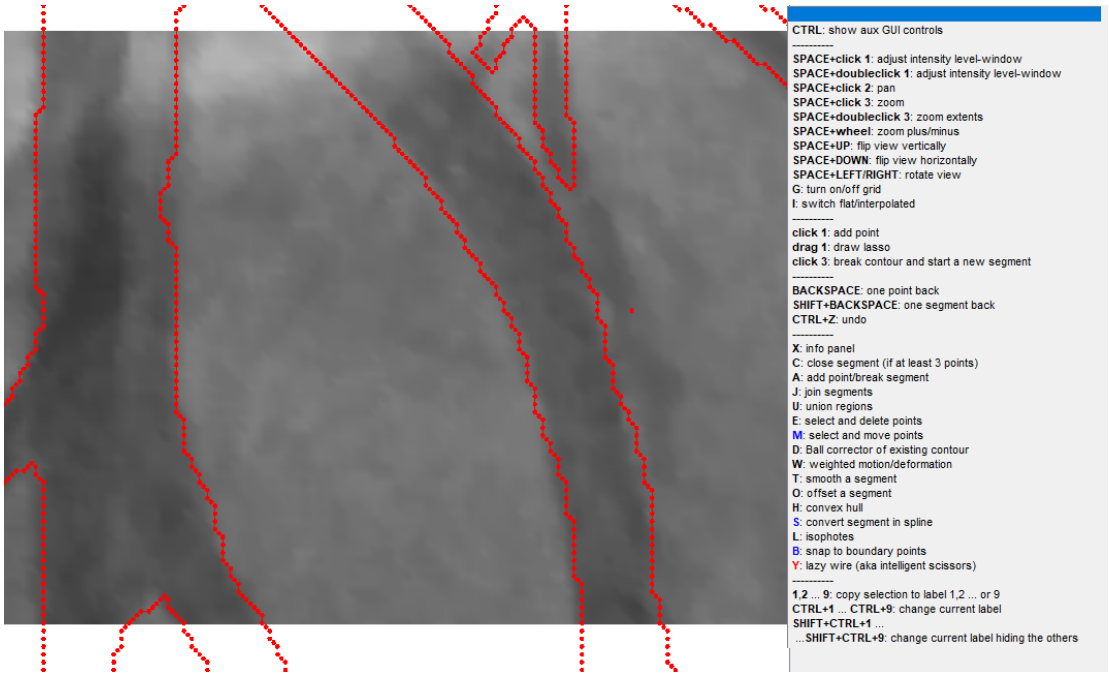


Figure 3.3: Matlab interface for vessel contour delineation. The red continuous points highlight the existing contours with the operation manual to the right.

for precise control or, for convenience, continuously draw red contour points by hand. To enhance efficiency, a rapid annotation feature allows users to select two points and automatically generate a connecting spline line. By joining different spline lines, users can create complete contours. For modifying existing contours, users can select red points to remove or reposition them. Additionally, a ball corrector with an adjustable diameter enables users to push or pull surrounding red points for fine adjustments. To ensure accurate contouring, the tool only allows saving if the segment forms a fully closed contour. It automatically identifies and connects any disconnected points to close the contour. To further assist users, the tool supports navigation through frames in the ICA sequence, making it easier to identify vessels that may be difficult to discern in the current frame.

3.3 SJTU Dataset D_2

The SJTU dataset D_2 originates from [34] and was acquired from the Renji Hospital of SJTU. The raw data consists of 120 sequences, each containing between 30 and 140 frames. These ICA sequences in the dataset are acquired from different machines (i.e., 800 mAh digital silhouette angiography X-ray machine from Siemens, medical angiography X-ray

system from Philips). They were resized to a resolution of 512×512 and processed using Poisson denoising [187] to account for variations caused by different acquisition machines. From these 120 sequences, 323 short sequences were extracted, each comprising four frames. For binary annotation, the third frame of each short sequence was selected as the target frame, resulting in a total of 323 annotated frames, with 269 frames corresponding to LAD and left circumflex artery (LCX), and 54 frames corresponding to RCA. Three experts conducted the annotation procedure, and the final ground truth was derived by averaging their annotations.However, the available dataset is provided in PNG format, which lacks essential metadata such as acquisition angles and original spatial resolution.

The general protocols for short sequence selection from the complete sequence, target frame selection, and annotation were not articulated in [34]. However, a careful evaluation of the dataset annotations reveals relatively low-quality manual segmentation. Examples of annotations for RCA and LAD, along with their corresponding ICA frames, are presented in Fig. 3.4. The figure highlights three major issues with dataset D_2 using arrows of three different colors. Thin vessels and those in low-contrast areas are often neglected, as indicated by the red arrows. The distal segments of the vascular tree are frequently short in length and exhibit boundary over-segmentation, as shown by the green arrows. Additionally, some segmented vessels fail to represent a complete vascular structure, with the segments in overlapped areas being incorrectly identified, as marked by the blue arrows.

3.4 Conclusion

The JR dataset D_1 and the SJTU dataset D_2 together serve as the foundation for the datasets used in this thesis to address different research objectives. However, the quality of segmentation between these two datasets varies significantly. The JR dataset D_1 provides a comprehensive vascular structure, neglecting only extra-thin vessels, while the SJTU dataset D_2 has several issues, including vessels missed in low-contrast areas, over-segmentation, and incorrect annotations in overlapping vessel regions. This distinction classifies D_2 as a more “coarse-grained” dataset and D_1 as a more “fine-grained” dataset.

The primary advantage of a “coarse-grained” dataset like D_2 lies in annotation ef-

ficiency: annotating an ICA at the quality level of D_1 typically takes about 1.5 hours, whereas annotating at the D_2 level takes roughly 45 minutes. However, the limitations in D_2 are expected to impact the training of architectures. The expert selected a subset of 10 ICA images from D_2 and re-segmented to reach the "fine-grained" quality to evaluate this impact quantitatively and qualitatively. Further details on this evaluation can be found in Section 5.3.8.

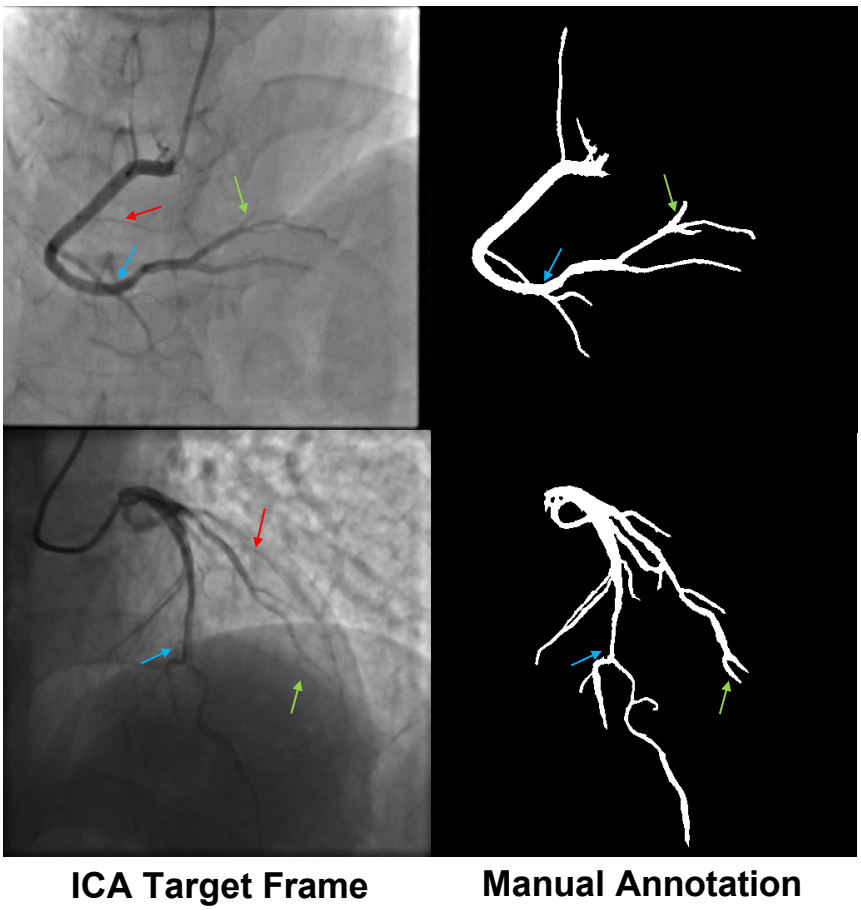


Figure 3.4: Annotated ICA frames and corresponding vessel masks for RCA (top) and LAD (bottom). Red arrows highlight the missed vessels. Green arrows highlight the over-segmented vessels. Blue arrows highlight the wrongly tracked vessels.

Chapter 4

Semi-Supervised ICA Segmentation

Chapter contents

Part of this chapter was presented in the paper "Semi-supervised coronary vessels segmentation from invasive coronary angiography with connectivity-preserving loss function," in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5, IEEE, 2022.

4.1 Introduction

In this chapter, I present a novel semi-supervised framework designed for segmenting invasive coronary angiography (ICA) images. This framework is inspired by the mean teacher model, which comprises two primary components: a student model and a teacher model [122]. By utilizing both labeled and unlabeled data, the framework effectively addresses the challenge posed by limited annotated datasets, a prevalent obstacle in multiple medical image analysis tasks, and, in particular, in ICA segmentation.

My approach optimizes segmentation performance by integrating supervised and unsupervised losses, enabling effective learning from both labeled and unlabeled inputs. Unlabeled inputs play a pivotal role in the process, serving three distinct purposes: inputs for the student model, inputs for the teacher model, and duplicated inputs for uncertainty estimation. This multi-faceted use of unlabeled inputs allows the framework to derive uncertainty insights from the teacher model, which then guides the unsupervised loss and enhances the learning of the student model. The student model's weights are transferred to the teacher model through iterative optimization, which is subsequently updated. For inference, the teacher model is retained as the final model, ensuring robust performance.

The backbone architecture of this framework is a modified Nested U-Net (UNet++), which excels at capturing multi-scale features and optimizing network depth [88]. UNet++ is enhanced with dropout layers to improve generalization and robustness. Additionally, I employ a connectivity-aware loss function, the elastic interaction energy loss, designed to mitigate segmentation discontinuities [188]. Traditional loss functions, such as binary cross-entropy and Dice loss, which aim to optimize pixel-wise accuracy or maximize the Dice coefficient, often struggle to preserve the continuity of vascular structures. This can result in fragmented or disconnected vessel branches. My approach ensures a more complete and accurate representation of the vessel tree, which is essential for downstream

applications.

This framework achieves superior segmentation outcomes by leveraging the diverse characteristics of unlabeled data and optimizing supervised and unsupervised losses. It demonstrates a comprehensive utilization of available data resources, significantly enhancing the segmentation quality even with limited ground truth.

Contributions

My contributions in this work can be summarized as:

1. The introduction of a novel semi-supervised framework for ICA segmentation that leverages both labeled and unlabeled data. This framework integrates a teacher-student model designed to optimize segmentation quality by minimizing supervised and unsupervised losses, guided by uncertainty estimation.

2. This method utilizes a UNet++ backbone with enhanced skip connections, enabling the extraction of robust multi-scale features while maintaining structural consistency. This architecture ensures efficient network training and supports model pruning for practical applications.

3. A connectivity-preserving loss function is incorporated to address the challenge of maintaining vascular tree integrity. This loss function reduces segmentation discontinuities and ensures topological accuracy, enabling the generation of reliable vessel skeletons.

4.2 Materials and Methods

In this section, I present the dataset and provide a comprehensive description of my mean teacher framework, which utilizes UNet++ as its backbone. I define I as the input ICA images and G as the corresponding gold standard, where $I \in [0, 255]^{H \times W}$ and $G \in \{0, 255\}^{H \times W}$ with H and W representing the height and width of the images. Consequently, I denote the labeled dataset as $D_L : \{(I_n, G_n)\}_{n=1}^{N}$ and the unlabeled input as $D_U : \{I_m\}_{m=1}^{M}$, where N represents the total number of labeled data and M denotes the number of unlabeled data instances. The segmentation processes conducted by the student model

and the teacher model are denoted as $f_s(\cdot)$ and $f_t(\cdot)$, respectively. The network weights and the noise are symbolized by θ and ζ, respectively.

4.2.1 Study Population

I perform the training and evaluation of the semi-supervised framework using dataset D_1, which consists of 120 ICA frames, including 73 right coronary artery (RCA) images and 47 left anterior descending artery (LAD) images. From the 60 labeled frames, a subset of 58 frames is randomly selected to form the labeled dataset (D_L), containing 27 RCA and 31 LAD images. The remaining 60 frames are used as the unlabeled dataset (D_U).

The labeled dataset D_L is further split randomly into three groups for training, validation, and testing. The training set comprises 42 frames (18 RCA and 24 LAD images), the validation set includes 6 frames (3 RCA and 3 LAD images), and the testing set contains 10 frames (6 RCA and 4 LAD images). This split ensures that the model's training set includes a diverse and representative subset of the labeled data (for RCA and LAD), while holding out separate data for unbiased performance evaluation.

4.2.2 UNet++

The UNet++ can be treated as an independent part of this framework, as it is still able to delineate vessels without other parts of the framework, given a large enough labeled set D_L. From the left of Fig. 4.1, the distinction between the UNet (depicted in black) and UNet++ lies in the addition of blue nodes and skip connections that bridge the gap between the encoder and decoder. In contrast to the original UNet, which utilizes a single skip connection to link the encoder and decoder at the same level, UNet++ employs multiple sub-UNets to establish these connections.

In the context of a pyramidal UNet++ structure with k levels, the total count of convolution nodes can be determined using the formula $[k \times (k+1)]/2$. Each convolution node is represented as $U^{i,j}$, with i representing the number of pooling or down-sampling operations from $U^{0,0}$ and j representing the number of received skip-connections of the node. Therefore, in a 5-level pyramidal structure, there are 15 $U^{i,j}$ blocks, where $i, j \in \{0, 1, 2, 3, 4\}$. The number of channels in the output of each node equals 2^{i+5}. With the six interlinked blue nodes between encoder and decoder ($U^{0,1}, U^{0,2}, U^{0,3}, U^{1,1}, U^{1,2}, U^{2,1}$), the connec-

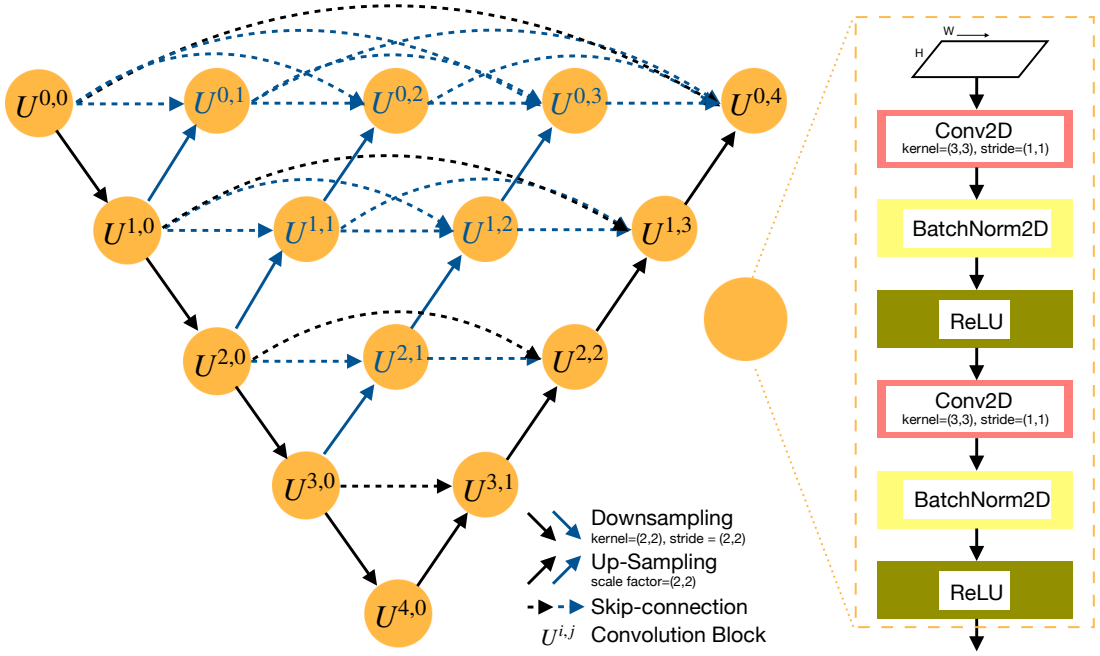


Figure 4.1: UNet++ structure with black paths and nodes representing a UNet of the same depth within a UNet++.

tion from $U^{0,0}$ to any node can be seen as a dense convolution block. For example, node $U^{1,2}$ receives two skip-connection outputs, and one up-sampling output concatenated together, where these outputs are generated in different convolution layers and pyramid levels. This narrowing of the semantic gap between the encoder and decoder facilitates the optimization process of the network and significantly encourages feature fusion.

$U^{i,j}$ encapsulate various convolutional operations. Specifically, I utilize a VGG convolution unit, which entails a sequence comprising convolution, batch normalization, and Rectified Linear Unit (ReLU) operations, iterated twice, as depicted on the right-hand side of Fig. 4.1. Mathematically, the operation inside UNet++ is expressed as:

$$U^{i,j} = \begin{cases} V(U^{i-1,j}) & j = 0 \\ V(J(\{U^{i,k}\}_{k=0}^{j-1}, R(U^{i+1,j-1}))) & j > 0 \end{cases} \tag{4.1}$$

where function $V(\cdot)$, $J(\cdot)$, and $R(\cdot)$ represent the convolution block, concatenation of features, and up-sampling from lower blocks, respectively. UNet++ is unique as every $U^{0,j}$ block can produce a supervised loss for pruning, which is discussed in later sections.

4.2.3 Mean Teacher Framework

The mean teacher model can be divided into student and teacher components, both adhering to an identical UNet++ architecture. Alternatively, the model can be partitioned into supervised and unsupervised components, reflecting the application of labeled input I_n and unlabeled input I_m, respectively. The supervised component solely resides within the student model, yielding the output $f_s(I_n)$. The unsupervised component entails outputs from both student and teacher models for unlabeled input ($f_s(I_m), f_t(I_m)$), accompanied by an uncertainty estimation method integrated with Monte Carlo Dropout [127], as depicted in Fig. 4.2. Although the quality of predictions on the unlabeled dataset is uncertain, consistency regularization suggests that even with minor perturbations added to the unlabeled dataset, the model can maintain stability and generate consistent outputs. Therefore, in the unsupervised component, noise is added to the unlabeled input I_m to promote greater consistency in the model's outputs.

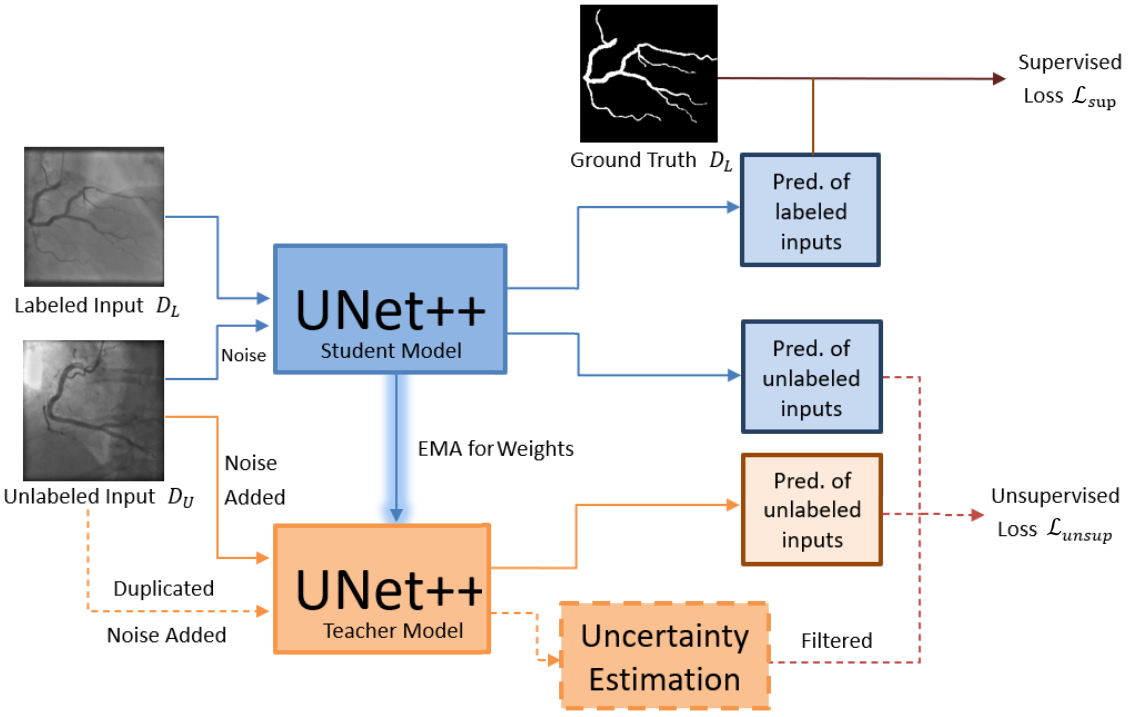


Figure 4.2: The framework for semi-supervised segmentation. The UNet++ blocks are just a simplification of the actual network, and they are identical in the teacher model and student model, with the student model in blue and the teacher model in orange.

The supervised component involves comparing G_n to $f_s(I_n; \theta_s)$ to obtain the su-

pervised loss $\mathcal{L}_{sup}$. Meanwhile, the unsupervised component comprises three distinct pathways. Firstly, the student model is trained using unlabeled I_m with Gaussian noise ζ_s, resulting in $f_s(I_m; \theta_s, \zeta_s)$. Subsequently, the same I_m, but with different Gaussian noise added, is inputted into the teacher model with random dropout applied before $U^{4,0}$ and $U^{0,4}$ (the bottom and last nodes of UNet++ in Fig. 4.1), producing $f_t(I_m; \theta_t, \zeta_t)$. Gaussian noise is used in this context because it is easy to generate and provides stable gradients for consistency training. Furthermore, since the selected ICA frames avoid those with low contrast-to-noise ratios, the noise characteristics in these frames allow the Poisson-distributed noise to be well approximated by Gaussian noise. Thus, the consistency distance between these two predictions is calculated as the image-scale mean squared error after applying softmax $S(\cdot)$ to both predictions:

$$Dist = \| S(f_s(I_m; \theta_s, \zeta_s), f_t(I_m; \theta_t, \zeta_t)) \|^2 \tag{4.2}$$

which has the same dimension as I_m. The most crucial step involves duplicating the same I_m for κ times to generate κ sets of predictions, each with different Gaussian noise added, thereby obtaining κ sets of predictions from the teacher model: $f_{t,r}(I_m; \theta_t, \zeta_{t,r})_{r=1}^{\kappa}$. These predictions are subjected to softmax transformation, and their averages across the total number of output channels yield a probability map μ of the same size as all predictions. This process enables the framework to assess the divergence of predictions under various perturbations, quantified as uncertainty u, calculated from predictive entropy due to its strictly bounded nature [127]. Mathematically, this uncertainty can be expressed as:

$$u = -\mu \cdot \log(\mu + \xi) \tag{4.3}$$

where ξ is a small value (1×10^{-6}) to avoid error at the start of the training. The uncertainty map u is generated at the image level, sharing the same dimensions as the gold standard. Conceptually, it functions akin to an "unlabeled" gold standard. Together with the distance metric $Dist$ obtained from Eq. (4.2), this uncertainty map facilitates the computation of the unsupervised loss $\mathcal{L}_{unsup}$.

Ensembling techniques [121, 122] play a pivotal role in enhancing the performance of semi-supervised models and reducing the variance between multiple versions of the teacher model. This is particularly crucial as weakly-supervised models can be prone

to bias towards the most recent training batches. While temporal ensembling [121] on predictions has been found inefficient, I adopted a similar ensemble approach on weights, inspired by [122]. Specifically, the weights of the student model θ_s are updated via standard gradient descent, whereas the weights of the teacher model are updated as an exponential moving average (EMA) of the student weights. Mathematically, the weight of the teacher model at the current iteration can be computed as follows:

$$\theta_t = \rho\theta_t^{'} + (1-\rho)\theta_s \tag{4.4}$$

where hyperparameter ρ governs the rate of EMA decay, and $\theta_t^{'}$ denotes the weights of the teacher model from the previous iteration.

4.2.4 Supervised Loss

The supervised loss function, inspired by the elastic energy of dislocations in crystals, leverages a level set method to represent topological changes. In the context of vascular structure segmentation, this effect can be visualized as depicted in Fig. 4.3(a), resembling an imaginary force acting between broken branches to mend vessel disconnections. I start by considering a single curve system. The energy system of this single curve $\gamma(x(s), y(s), z(s))$ consists of three key functions defined as [188]:

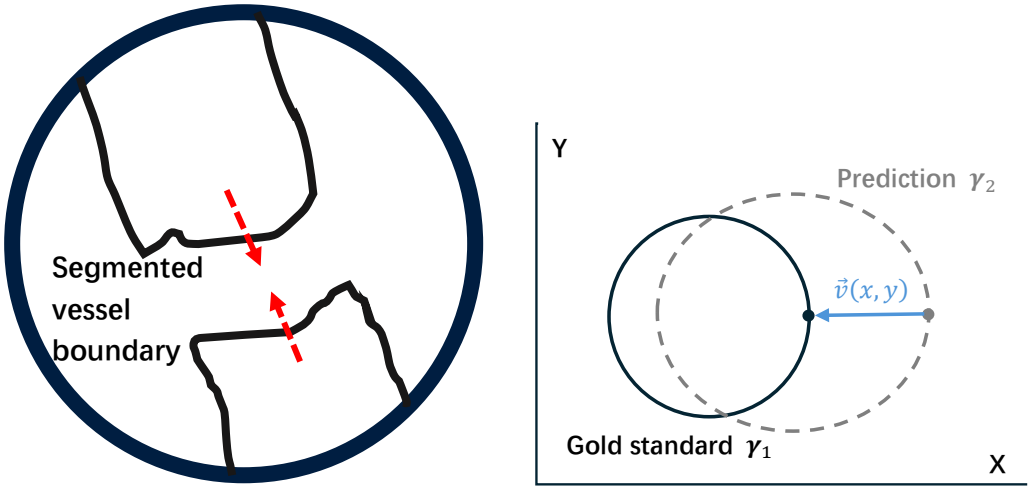
$$\nabla \times \vec{w} = \delta^2(\gamma)\vec{\tau} \tag{4.5}$$

$$\vec{w}(x,y,z) = -\frac{1}{4\pi}\int_{\gamma}\frac{\vec{r}\times d\vec{l}}{|\vec{r}|^3} \tag{4.6}$$

$$E = \frac{1}{8\pi}\int_{\gamma}\int_{\gamma'}\frac{d\vec{l}\cdot d\vec{l'}}{|\vec{r}|} \tag{4.7}$$

where $\delta^2(\gamma)$ is a delta function in 2D, $\vec{\tau}$ is the unit tangent vector of γ, $d\vec{l} = \vec{\tau}\delta^2(\gamma)dxdydz$ is an infinitely small part of the curve, $\vec{w} = (w_1(x,y,z), w_2(x,y,z), w_3(x,y,z))$ is a vector in xyz plane and $\vec{r} = (x - x(s), y - y(s), z - z(s))$ is a vector between points.

More specifically, Eq. (4.5) serves as the constraint function for the energy system, ensuring its consistency regardless of variations in the curve. Eq. (4.6) describes the system's dynamics under the specified constraint, while Eq. (4.7) defines the total elastic energy of the system, representing an element of the curve as $d\vec{l}$. Additionally, considering the properties of vectors $\vec{r}$ and $d\vec{l}$, $\vec{w}$ can be geometrically interpreted as a vector perpendicular to the plane where the curve is located.



(a) (b)

Figure 4.3: (a) Illustration of the effect of applying $\mathcal{L}_{sup}$ on a broken branch. (b) The 2D image plane contains two curves derived from the gold standard (stationary) and prediction (moving). The velocity of a point on the moving curve is illustrated using the blue arrow.

This system is not only valid on a single curve but also on a collection of curves. In Fig. 4.3(b), I define the stationary boundary of gold standard as γ_1 and the moving boundary of prediction as γ_2 in 2D image plane with $z = 0$. Subsequently, γ in previous functions from Eqs. (4.5) to (4.7) is substituted with $\gamma_1 \cup \gamma_2$ for applying this system on image segmentation. By introducing a moving curve, the velocity to describe its movement at time t has to be defined [188]:

$$\vec{v}(x,y) = \frac{\partial \gamma}{\partial t} = \vec{w} \times \vec{\tau} \tag{4.8}$$

In characterizing the evolving boundary, I employ the level set method. Any evolving curve can be conceptualized as a cross-section of a curved surface or level set ϕ, where the curve changes accordingly as the cross-sectional plane shifts. The velocity of this moving curve is defined implicitly in the evolution equation of the level set ϕ:

$$\frac{\partial \phi}{\partial t} = \vec{v}(x,y)|\nabla \phi| \tag{4.9}$$

which corresponds to the magnitude of the pointer in Fig. 4.3(a) and the moving curve is normally represented by the zero level set function, which is $\phi(x,y,0)$ in this case. This level set function is linked to prediction from the student model by $\phi(x,y,0) = 0.5 -$

$Softmax(f_s(I_n;\theta_s))$. Besides, I simply use G_s as a stationary boundary, which can be obtained by convolving the gold standard with a two-dimensional Gaussian function.

Therefore, the unit tangent vectors $\vec{\tau_1}$, $\vec{\tau_2}$ and the 2D delta function of stationary boundary γ_1 and the moving boundary γ_2 can be obtained separately as:

$$\vec{\tau_1} = \frac{\nabla G_s}{|\nabla G_s|} \times \vec{k} \quad \text{and} \quad \delta^2(\gamma_1) = |\nabla G_s|\delta(z) \tag{4.10}$$

$$\vec{\tau_2} = \frac{\nabla \phi}{|\nabla \phi|} \times \vec{k} \quad \text{and} \quad \delta^2(\gamma_2) = \delta^2(\phi)|\nabla \phi|\delta(z) \tag{4.11}$$

where $\delta(z)$ is the delta function of z and $\vec{k}$ is a unit vector along z-axis. As mentioned before, $\gamma = \{\gamma_1 \cup \gamma_2\}$, which enables us to linearly combine Eqs. (4.10) and (4.11) for $\delta^2(\gamma)\vec{\tau}$ in Eq. (4.5):

$$\nabla \times \vec{w} = \nabla(G_s + \alpha\eta_{step}(\phi)) \times \vec{k}\delta(z) = \delta^2(\gamma)\vec{\tau} \tag{4.12}$$

where $\eta_{step}(\phi) = \delta^2(\phi)\nabla\phi$ as $\eta_{step}(\cdot)$ is a regularized Heaviside step function and α is a hyperparameter to control the strength of dynamics in the moving boundaries. The regularized Heaviside function is accompanied by a regularization factor β:

$$\eta_{step}(\phi) = \begin{cases} 1 & \phi \leq -\beta \\ \frac{1}{2}(sin(1 + \frac{\pi\phi}{2\beta})) & -\beta < \phi < \beta \\ 0 & \phi \geq \beta \end{cases} \tag{4.13}$$

For calculation efficiency, I use a HardTanh function with $(min, max) = (0, 1)$ to approximate the normally used regularized Heaviside function for image smoothing [188]. Due to the property of $\vec{w}$ and the position of image plane ($z = 0$), it can be proved that $w_1 = w_2 = 0$. Therefore, Eq. (4.8) can be simplified with Eq. (4.11) as:

$$\vec{v}(x, y) = w_3\vec{k} \times \frac{\nabla\phi \times \vec{k}}{|\nabla\phi|} = \vec{w} \cdot \vec{k}\frac{\nabla\phi}{|\nabla\phi|} \tag{4.14}$$

Then substituting Eq. (4.12) into $d\vec{l}$ and Eq. (4.6) into Eq. (4.14), the velocity of moving boundary in Eq. (4.9) is calculated as:

$$v(x, y) = -\frac{1}{4\pi}\int_{\Re^2} \frac{\vec{r} \cdot \nabla(G_s + \alpha\eta_{step}(\phi))}{|\vec{r}|^3} dxdy \tag{4.15}$$

Similarly, the total energy can be derived from Eq. (4.7) and I define $G_s + \alpha\eta_{step}(\phi) = \mathbb{T}(x, y)$ for convenience:

$$E = \frac{1}{8\pi}\int_{\Re^3} dxdy \int_{\Re^3} \frac{\nabla\mathbb{T}(x, y) \cdot \nabla\mathbb{T}(x', y')}{|\vec{r}|} dx'dy' = \mathcal{L}_{sup} \tag{4.16}$$

Finally, Eq. (4.16) is the supervised loss function $\mathcal{L}_{sup}$ for this energy system, and Eq. (4.15) is the corresponding gradient of the loss function. To boost the efficiency during backpropagation, Eqs. (4.15) and (4.16) can be further simplified by transforming them into 2D Fourier space.

4.2.5 Unsupervised Loss

For the unsupervised loss, the approach involves filtering and retaining the uncertainty map u from Eq. (4.3), keeping only values smaller than a threshold T_{sig} to generate a mask. This pixel-wise operation ensures that only reliable features, characterized by relatively low uncertainty (given that high uncertainty correlates with low accuracy), are preserved. The threshold ϑ plays a critical role in determining the extent to which the uncertainty map is filtered. I define it as a sigmoid ramp-up function of the uncertainty:

$$\vartheta = 0.5u_{max}(1 + e^{(-5(1-\frac{Iter}{Iter_{max}})^2)}) \tag{4.17}$$

where u_{max} is the maximum uncertainty value, $Iter$ is the current number of iterations and $Iter_{max}$ is the maximum number of iterations in training. Such a threshold gradually filters out fewer uncertainty values as it increases, reflecting the fact that the teacher model is nearly random in earlier iterations but becomes more reliable in later iterations.

The loss value is then calculated by summing up all the confidence values within the mask for the consistency distance $Dist$ and dividing it by the total remaining uncertainty. The unsupervised loss function can be summarized as:

$$\mathcal{L}_{unsup} = \frac{\Sigma(u < \vartheta) \cdot Dist}{2 \cdot \Sigma(u < \vartheta) + \xi} \tag{4.18}$$

where the summation is pixel-wise to get a single value with ξ for avoiding error at the start of the training. Since the optimization only applies to θ_s and the loss function is entirely linear concerning $f_s(I_m; \theta_s, \zeta_s)$, the gradient can be automatically calculated and back-propagated.

Hence, the objective of this semi-supervised framework for ICA segmentation is to minimize the weighted sum of two loss functions:

$$\min_{\theta_s}[\mathcal{L}_{sup}(f_s, G_n) + \lambda_{unsup}\mathcal{L}_{unsup}(f_s, f_t)] \tag{4.19}$$

where λ_{unsup} is the hyperparameter that controls the balance between supervised and unsupervised loss.

4.2.6 Evaluation Methods

I employ two conventional metrics alongside three additional metrics to assess segmentation quality. The conventional metrics focus on Dice and recall, while the other metrics include over-segmentation ($\mathcal{H}(f|G)$), under-segmentation ($\mathcal{H}(G|f)$), and Betti number error. The Dice coefficient quantifies the degree of overlap between the predicted segmentation and the gold standard, while recall evaluates the model's ability to classify all vessel pixels present in the gold standard correctly. They are defined as follows:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{4.20}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.21}$$

where true positive (TP) is the total number of correctly classified pixels in vessel regions of the predicted vessel segmentation, false positive (FP) represent background pixels that are incorrectly classified as vessels, true negative (TN) represent background pixels correctly predicted as such, and false negative (FN) represent vessel pixels that were mistakenly labeled as background in the prediction.

The sum of over-segmentation and under-segmentation yields the Variation of Information (VOI), a measure of the distance between two segmentations. Quantitatively, the conditional entropies corresponding to over-segmentation and under-segmentation are computed as:

$$\mathcal{H}(f \mid G) = -\sum_{g \in G} P(g) \sum_{f' \in f} P(f' \mid g) \log P(f' \mid g) \tag{4.22}$$

$$\mathcal{H}(G \mid f) = -\sum_{f' \in f} P(f') \sum_{g \in G} P(g \mid f') \log P(g \mid f') \tag{4.23}$$

where P represents probability distributions based on pixel-wise co-occurrence.

Betti number error specifically evaluates the effectiveness of the elastic energy loss function introduced previously by comparing the topology (number of connected components) of 2-dimensional subjects. It is defined as:

$$\text{Betti} = \left| \mathcal{B}^{\text{Pred}} - \mathcal{B}^{\text{GS}} \right|, \tag{4.24}$$

where $\mathcal{B}^{\text{Pred}}$ and $\mathcal{B}^{\text{GS}}$ denote the number of connected components in the predicted and gold standard segmentations, respectively.

4.3 Experimental Results

4.3.1 Experimental Settings

This framework encompasses numerous settings and hyperparameters, many of which cannot be directly assigned fixed values. For instance, the weight of the unsupervised loss function in Eq. (4.19) is determined as $\lambda_{unsup} = 0.1 \cdot e^{(-5(1-\frac{Iter}{Iter_{max}})^2)}$, similar to the ramp-up for ϑ. Initially, λ_{unsup} is minimal, gradually increasing over training epochs. This design compels the network to rely predominantly on supervised loss in the early stages when the teacher model's predictions are arbitrary. Subsequently, as predictions from the teacher model become reliable, the semi-supervised model is gradually activated. Additionally, after performing a random grid search, I set the hyperparameter ρ in Eq. (4.4), which controls the update of weights in the teacher model, to 0.99, and α in Eq. (4.12), which regulates the dynamics of vessel boundary in predictions, to 0.3.

The network training is conducted on an NVIDIA V100 Tensor Core GPU utilizing the Pytorch package. The hyperparameter, κ, for uncertainty estimation is set to 10, constrained by GPU memory limitations, and dropout is exclusively activated during training. I optimize the network using Stochastic Gradient Descent (SGD) with a weight decay of 1×10^{-6} and momentum of 0.7. Initially, the learning rate is set to 1×10^{-5} but is modulated by a cosine annealing scheduler, gradually decreasing to its minimum value of 1×10^{-8} throughout epochs, following a cosine curve pattern. Training persists for 10,000 iterations to ensure convergence of the loss function. Each batch comprises 10 samples, including 4 labeled and 6 unlabeled images. On-the-fly augmentation is employed, incorporating standard random rotation, flipping, saturation, and contrast adjustments to introduce diversity and stochasticity. The size of convolutional operations is fixed at 3×3. The model selected for testing is the one that attains the highest Dice score on the validation dataset during training.

4.3.2 Model Pruning Analysis

Training CNNs is resource-intensive and time-consuming, particularly for complex semi-supervised frameworks. While deeper networks may offer increased capacity, their performance benefits are not always proportional to the computational costs involved. Therefore, it is imperative to ascertain the optimal depth of the UNet++ architecture that balances performance gains with computational efficiency, considering the limitations of GPU memory, for further implementation in the mean teacher model.

Before proceeding with other experiments, I conducted four preliminary trials using different depths of UNet++. Denoted as L_d, each trial involved transforming UNet++ L_d to UNet++ L_{d-1} by removing all nodes in the decoder of UNet++ L_d. Trained for 6000 iterations, a model was saved at the end of each trial and evaluated on the same test set. The time taken for testing was recorded for each model, as illustrated in Fig. 4.4.

It can be observed that UNet++ L_1 exhibits the largest performance gap compared to other networks, apart from the degree of over-segmentation in Fig. 4.4(d), which indicates that aggressive pruning is not applicable in this case. The least difference happens between UNet++ L_2 and L_3 where Dice only improves by 0.87%. The most substantial percentage difference observed by transforming UNet++ L_d to UNet++ L_{d-1} is in the Betti number, with a notable 59.94% improvement from UNet++ L_1 to L_2. Besides, the network's performance improves when the depth is changed from 2 to 4, demonstrating an exponential trend, albeit not as significant as the substantial boost between UNet++ L_1 and L_2. Considering UNet++ L_4 still performs best within memory limits, I decided not to prune the network during subsequent training. However, for a more convenient application of this framework, the network could be pruned to UNet++ L_3, as it offers 17.72% less inference time for only 1.11% less Dice and 1.30% less recall.

In conclusion, while the performance gain from UNet++ L_2 to L_3 is modest, L_3 offers a more efficient alternative to L_4, making it preferable when computational resources are limited. In contrast, L_4 remains the best choice when accuracy is the highest priority.

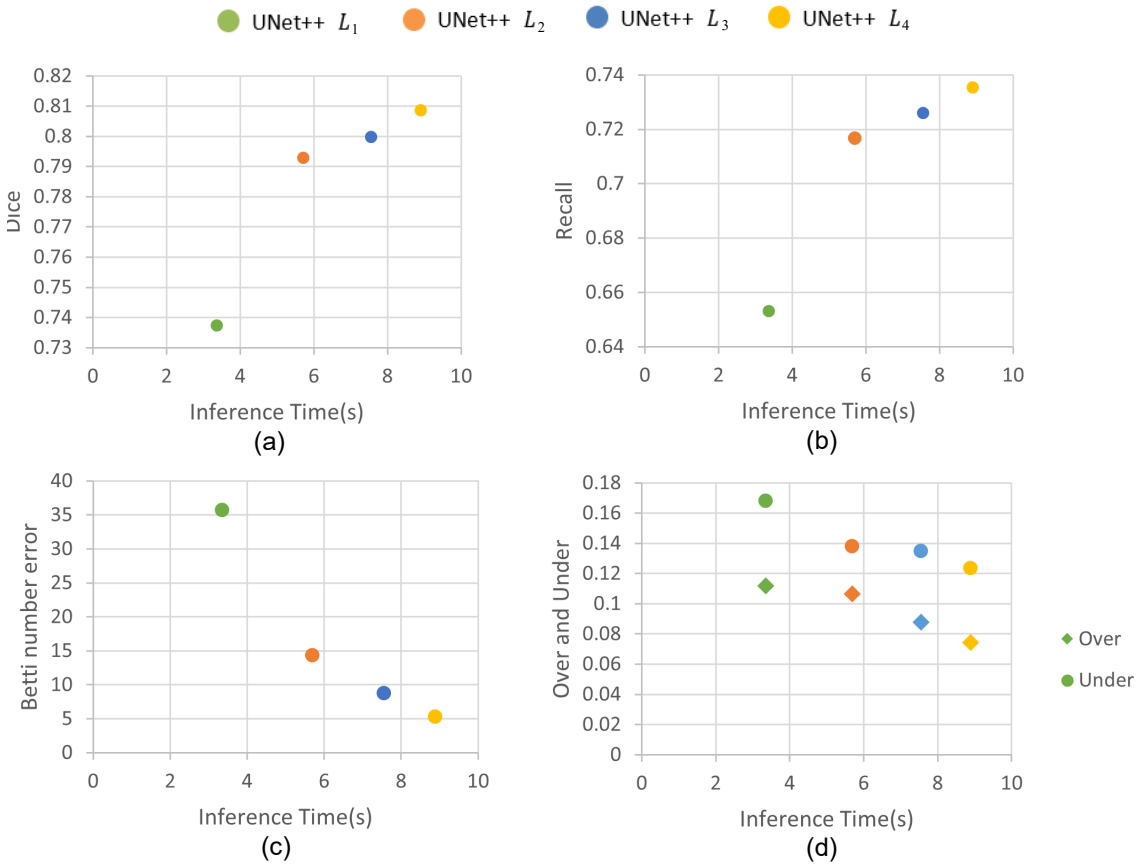


Figure 4.4: Evaluation metrics including Dice coefficient (a), recall (b), Betti number error (c), and over- and under-segmentation (d) are depicted with scatter points in this figure. The same metric from UNet++ with different pyramidal levels is compared in the same sub-figure with different colors to represent network depth of 1, 2, 3, and 4.

4.3.3 Impact of Supervised Loss Functions: Dice Loss vs. $\mathcal{L}_{sup}$

The choice of supervised loss functions significantly influences segmentation performance. I compare the widely used binary cross-entropy Dice loss (Dice) and the elastic energy-related loss ($\mathcal{L}_{sup}$) across all frameworks. The Dice loss, defined as $0.5 \times (\text{Binary Cross Entropy Loss} + \text{Dice Loss})$, offers a balance between accuracy and computational efficiency but struggles with maintaining vessel continuity, especially at bifurcations and branch points. As shown in Fig. 4.5, frameworks employing Dice loss exhibit higher levels of noise, with more false positives (blue areas) and false negatives (red areas), complicating post-processing steps like connected component analysis.

In contrast, $\mathcal{L}_{sup}$ is designed to prioritize structural continuity and penalize disconnected segments, leading to significant improvements in vessel connectivity. This is quan-

Table 4.1: Comparison between the proposed method and state-of-the-art approaches with increasing framework complexity. All metrics are calculated after applying a mean threshold on predictions to identify the intensity of background for binarization. N and M represent the number of labeled and unlabeled images, respectively.

Method	Loss	Dataset (N/M)	Dice (%)	Recall (%)	Over (%)	Under (%)	Betti
UNet	Dice	42 / 0	77.34±2.42	79.05±5.95	14.48±3.77	13.93±3.30	22.76±10.72
UNet	$\mathcal{L}_{sup}$	42 / 0	79.45±2.01	80.12±5.97	13.15±3.64	11.29±2.38	8.25±4.57
UNet++	Dice	42 / 0	78.84±1.83	80.23±5.67	14.38±3.68	14.15±3.14	27.89±12.72
UNet++	$\mathcal{L}_{sup}$	42 / 0	80.86±3.63	81.27±6.33	**11.01±2.99**	11.67±2.43	7.28±4.22
MT-UNet	Dice	42 / 60	76.81±2.24	81.89±5.27	14.83±2.86	12.03±2.77	30.8±12.27
MT-UNet	$\mathcal{L}_{sup}$	42 / 60	78.59±2.64	81.48±5.42	13.40±2.03	11.64±2.42	7.15±4.81
MT-UNet++	Dice	42 / 60	79.52±1.97	82.06±5.16	13.09±2.54	11.20±2.49	20.48±6.68
MT-UNet++	$\mathcal{L}_{sup}$	42 / 60	**81.66±1.89**	**83.89±4.81**	11.19±2.33	**10.62±1.84**	**6.83±3.87**
SS-CADA [189]		60 / 92	78.4±4.60	83.27±3.95	N/A	N/A	N/A

All results represent mean ± standard deviation (SD).

titatively reflected in Table 4.1, where frameworks using $\mathcal{L}_{sup}$ consistently achieve lower Betti numbers, indicating fewer segmentation discontinuities. For instance, MT-UNet++ ($\mathcal{L}_{sup}$) achieves a Betti number of 6.83, the smallest among all frameworks. Furthermore, $\mathcal{L}_{sup}$ reduces over-segmentation (false positives) within testing groups by incorporating elastic interaction principles, as evidenced by the smaller blue areas in Fig. 4.5 and the zoomed views in Fig. 4.6. However, a trade-off is observed: $\mathcal{L}_{sup}$ slightly reduces recall compared to Dice loss, as the latter's tendency to segment thicker vessels captures a larger positive area, resulting in a peak recall of 81.89% for MT-UNet (Dice).

4.3.4 Comparison of Network Architectures: UNet++ vs. UNet

The architectural differences between UNet and UNet++ play a critical role in segmentation performance, particularly for complex vascular structures. UNet++ introduces a nested structure with dense skip connections and multi-scale feature fusion, addressing the limitations of UNet in delineating vessels with varying diameters and intricate branching patterns. As shown in Table 4.1, replacing UNet with UNet++ yields performance gains across all metrics. For example, UNet++ ($\mathcal{L}_{sup}$) achieves a Dice coefficient of 80.86%, compared to 79.45% for UNet ($\mathcal{L}_{sup}$), an improvement of approximately 1.77%. Similarly, over-segmentation is markedly reduced, with UNet++ ($\mathcal{L}_{sup}$) achieving an over-segmentation rate of 0.1101, compared to 0.1315 for UNet ($\mathcal{L}_{sup}$), representing a 19.43% reduction.

The enhanced connectivity and feature fusion capabilities of UNet++ are particularly beneficial for preserving vessel structures, as highlighted in Fig. 4.5 and the zoomed views

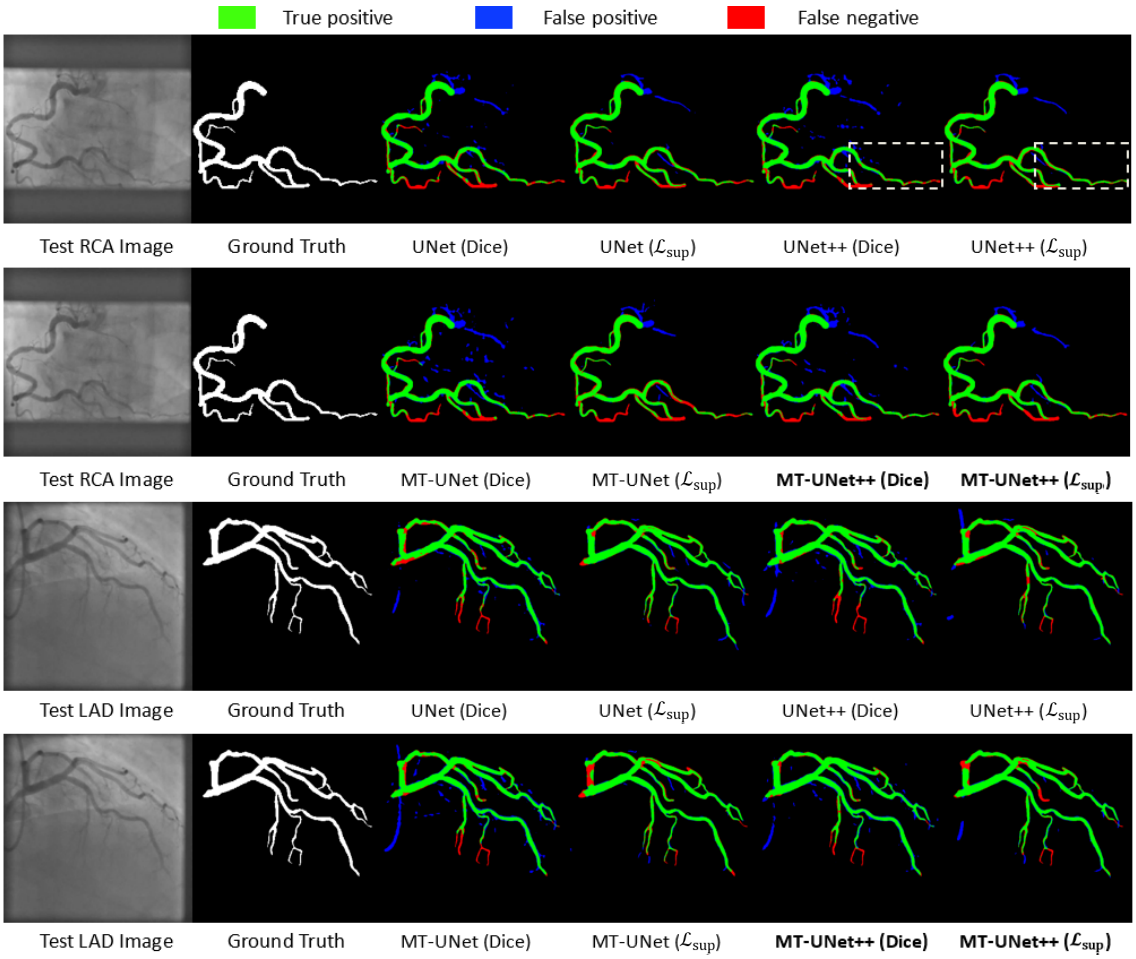


Figure 4.5: Two sets of segmentation test results with one RCA image and one LAD image. The names of 8 different frameworks in Table 4.1 with binary cross entropy Dice loss (Dice) and the elastic energy related loss ($\mathcal{L}_{sup}$) are labeled under corresponding results. True positive (TP) is in green, while false positive (FP) and false negative (FN) are in blue and red. The first two rows correspond to the first set, and the last two rows correspond to the second set. The dotted boxes are regions for zooming.

in Fig. 4.6. The dense skip connections allow for effective multi-scale representation, which is critical for handling the wide range of vessel sizes in ICA images. These advantages make UNet++ a superior choice for frameworks that prioritize segmentation quality and vessel connectivity.

4.3.5 Effectiveness of the Mean Teacher Framework

The incorporation of the mean teacher framework, which leverages consistency loss and uncertainty estimation, provides substantial benefits for semi-supervised segmentation. By utilizing unlabeled data alongside labeled data, the framework enhances the model's

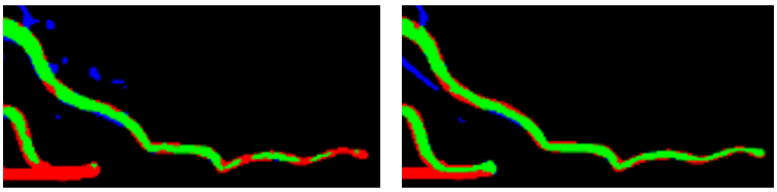


Figure 4.6: The zoomed views in the dotted boxes in Fig. 4.5, illustrating improvements from elastic energy loss.

ability to generalize and segment underrepresented structures, such as thin vessels. Table 4.1 highlights the impact of the mean teacher framework, with reductions in under-segmentation (red areas in Fig. 4.5) of 9.60% for MT-UNet++ ($\mathcal{L}_{sup}$) compared to UNet++ ($\mathcal{L}_{sup}$).

Despite these improvements, the mean teacher framework tends to slightly increase over-segmentation, as it captures thin, vessel-like structures that may not be included in the gold standard annotations. This effect is visible in Fig. 4.5, where the additional positives introduced by the mean teacher often align with actual vessels, enhancing its robustness to weakly-labeled data. Furthermore, the mean teacher framework contributes to a consistent increase in recall, with a mean improvement of approximately 2.28% for MT-UNet++ (Dice) over UNet++ (Dice). My work also compares favourably with the previous state-of-the-art method, Semi-Supervised Cross-Anatomy Domain Adaptation (SS-CADA) [189], with improvements of 4.15% and 0.78% in Dice and recall, respectively.

Overall, the mean teacher framework complements the strengths of UNet++ and $\mathcal{L}_{sup}$, achieving the highest Dice coefficient of 81.66% for MT-UNet++ ($\mathcal{L}_{sup}$). These results underscore the effectiveness of combining semi-supervised learning with advanced architectures and connectivity-aware loss functions for ICA segmentation.

4.3.6 Impact of the Number of Labeled Samples

The semi-supervised framework's superiority over the supervised framework is expected to vary within a range of labeled samples. Thus, a comprehensive ablation study examining performance across different numbers of annotated samples is essential. Focusing on MT-UNet++ ($\mathcal{L}_{sup}$) and UNet++ ($\mathcal{L}_{sup}$), I conducted 19 training iterations with varying numbers of labeled data $N = 4, 6, 8, ..., 38, 40$ while maintaining consistent setups. The

results of the final inference metrics are depicted in Fig. 4.7. The initial value of the number of labeled samples is set at 2, excluding $N = 2$ for UNet++ ($\mathcal{L}_{sup}$) due to its inability to segment anything, resulting in empty prediction images.

In general, four types of outcomes are possible when an evaluation metric is plotted against the number of labeled data from both a semi-supervised framework and a supervised framework. Denoting ψ_1 and ψ_2 as metrics evaluated on the semi-supervised and supervised frameworks, respectively, these outcomes are:

1. $\psi_1 > \psi_2$ when N is small, but $\psi_1 < \psi_2$ after N increases
2. $\psi_1 < \psi_2$ when N is small, but $\psi_1 > \psi_2$ after N increases
3. $\psi_1 > \psi_2$ for all available N
4. $\psi_1 < \psi_2$ for all available N

ψ_1 and ψ_2 variations are depicted as line charts in Figs. 4.7(a-e). However, due to the randomness inherent in neural networks and the relatively small step size on the x-axis, discerning the trend of metrics directly can be challenging. Therefore, a fifth-order polynomial trend line is added as a dotted curve to approximate each solid actual curve (line). Figs. 4.7(b,c) represent the third type of outcome, while Fig. 4.7(d) embodies the fourth type. For instance, Fig. 4.7(b) illustrates the chart for recall, consistently showing the enhancement in recall with the introduction of the mean teacher semi-supervised framework, corroborating previous analyses. Similarly, in Fig. 4.7(c) and Fig. 4.7(d), the mean teacher model exhibits superior performance in minimizing under-segmentation but tends to over-segment thin vessels influenced by the teacher model. Conversely, Figs. 4.7(a,e) represent approximately the first type of outcome, but with an opposite interpretation. Specifically, Fig. 4.7(a) demonstrates the chart for the Dice coefficient, revealing that when the total number of labeled data $N < 22$, the semi-supervised framework outperforms the supervised network. However, when $N > 22$, the difference between the two frameworks diminishes. Betti number error in Fig. 4.7(e) serves mainly to illustrate the influence of $\mathcal{L}_{sup}$ and is the least significant metric in this study. Consequently, it can be inferred that the optimal operating range for the semi-supervised framework is when $N < 22$, with the

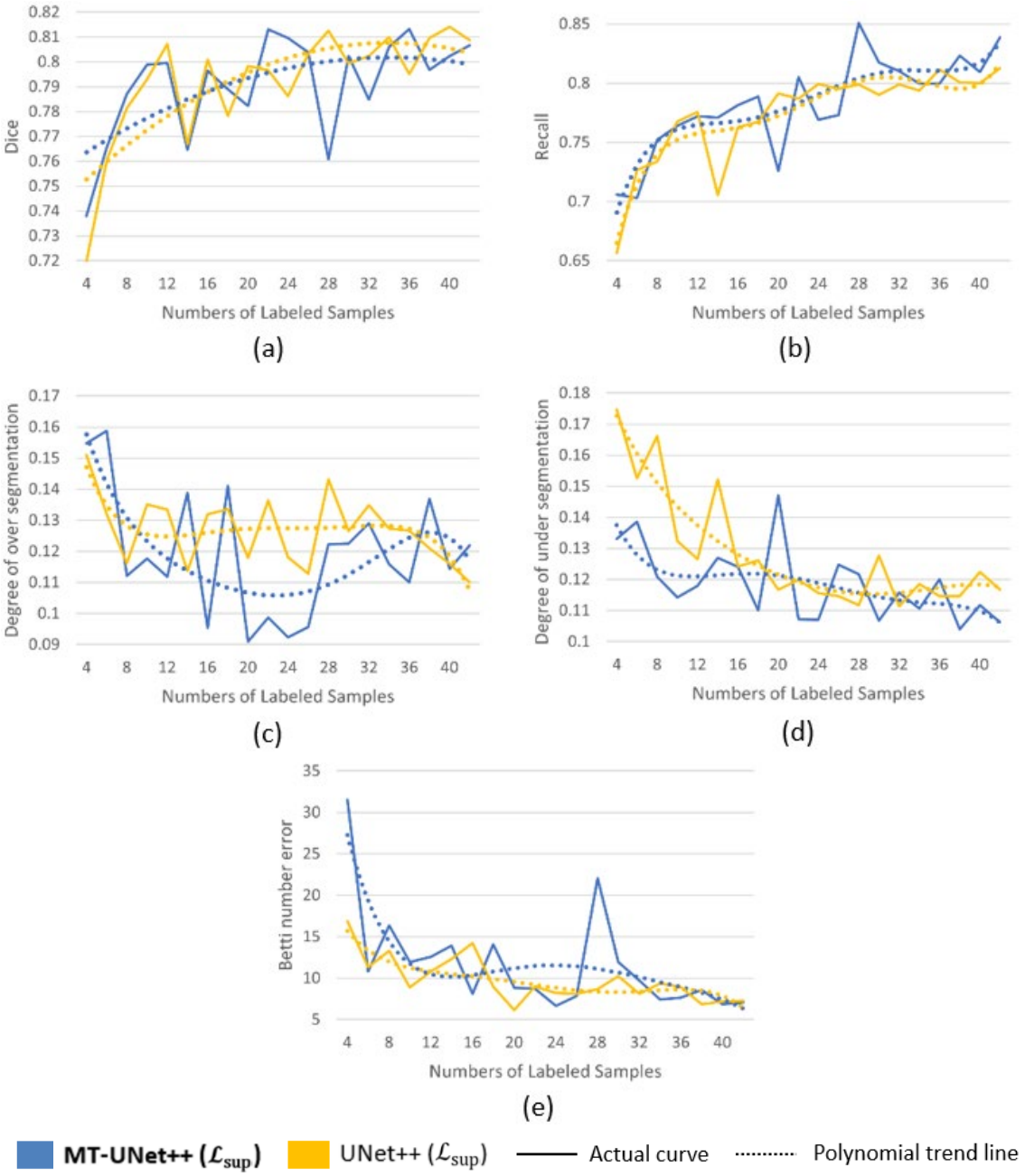


Figure 4.7: Dice (a), recall (b), over-segmentation (c), under-segmentation (d) and Betti number error (e) of MT-UNet++ ($\mathcal{L}_{sup}$) and UNet++ ($\mathcal{L}_{sup}$) with respect to the number of labeled samples during training.

overall performance evaluated by the Dice coefficient in Fig. 4.7(a), as the relationship of other metrics remains consistent across all values of N.

4.4 Discussion and Conclusion

This study presents an innovative semi-supervised framework for segmenting ICA images, tackling the issues of restricted labeled datasets and the complexity of vascular segmentation. Utilizing the mean teacher model, the framework merges supervised and unsupervised learning by incorporating both labeled and unlabeled data. The Nested UNet (UNet++) architecture plays a key role in achieving strong segmentation results. Its dense skip connections enhance multi-scale feature extraction, which is essential for understanding the intricate structure of vascular networks. Additionally, its pruning capabilities contribute to computational efficiency, allowing the framework to adapt to various hardware limitations. The pruning results suggest that UNet++ L_3 meets several key performance metrics. This makes it suitable for practical applications when computational resources are constrained. Such cases may include adapting to different annotation protocols or changes in ICA machine settings. UNet++ L_3 enables straightforward implementation and sets a benchmark for future improvements. By incorporating supervised and consistency losses, this approach enables the student model to leverage uncertainty-guided insights from the teacher model, resulting in improved performance even with a limited number of annotations.

The experimental results demonstrated the proposed framework's superiority over other methods. It achieved the highest Dice coefficients of 81.66% and recall of 83.89% while minimizing over- and under-segmentation. The elastic interaction energy loss not only improved connectivity but also reduced segmentation noise, ensuring a cleaner and more precise delineation of the vascular structures. Notably, the semi-supervised approach excelled in extracting valuable information from unlabeled data, effectively augmenting the learning process of the student model and significantly improving segmentation quality on labeled datasets.

Furthermore, the ablation study indicates that the mean teacher semi-supervised framework demonstrates a superior performance advantage with a reduced number of labeled samples, especially when utilizing fewer than 22 samples in this experiment. The advantage becomes increasingly pronounced as the number of labeled samples diminishes. However, this quantity is highly specific to the task, as varying annotation

protocols (such as single main vessel segmentation versus full vascular tree segmentation) and the selection of the target frame within the cardiac cycle, along with other preliminary setups, can influence this figure.

In summary, the proposed framework establishes a robust and effective approach for ICA segmentation, addressing critical limitations in existing methodologies. By combining innovative architectural designs and advanced loss functions within a semi-supervised paradigm, the method achieves SOTA performance, offering a practical solution for clinical applications where labeled data is often scarce.

Chapter 5

Temporal Vessels Segmentation

Chapter contents

Part of this chapter was presented in the journal "Deep learning based coronary vessels segmentation in X-ray angiography using temporal information," Medical Image Analysis, vol. 102, p. 103496, 2025.

5.1 Introduction

The automated segmentation of coronary vessels in invasive coronary angiography (ICA) poses significant challenges, even when ample data is accessible. Although challenges such as motion artifacts and uneven contrast distribution are inherent to ICA imaging, the availability of a larger annotated dataset provides an opportunity to develop and evaluate more robust methods to mitigate these issues. In Chapter 4, I addressed the issue of limited data availability by employing a semi-supervised learning framework, which utilized both labeled and unlabeled samples to attain robust segmentation. Nevertheless, this semi-supervised methodology is limited by the intrinsic constraints associated with sparse annotations and is unable to entirely resolve the complex dynamics evident in ICA sequences.

In this chapter, attention is directed towards scenarios in which a sufficient volume of annotated ICA data is accessible. With an expanded dataset, the challenges transcend mere data scarcity, encompassing the attainment of high fidelity in segmentation through the comprehensive utilization of the temporal information embedded within ICA sequences. Motion artifacts, overlapping structures, and varying diameters of vessels persist as significant issues; however, the provision of temporal information in multiple frames presents an opportunity to alleviate these challenges. By capturing spatial-temporal relationships across successive frames, segmentation can be refined not only in terms of accuracy but also in the preservation of the structural integrity of the vascular architecture, particularly concerning thin vessels.

To this end, I propose the Temporal Vessel Segmentation Networks (TVS-Net and TVS-Net+), two novel architectures designed to leverage the temporal dynamics in ICA sequences. Unlike traditional single-frame segmentation methods, TVS-Net extracts

spatial-temporal features by incorporating time as an additional dimension. The architecture employs a densely nested three-dimensional (3D) encoder for temporal feature extraction and a highly connected two-dimensional (2D) decoder for precise spatial segmentation, ensuring that information from consecutive frames is seamlessly fused to improve performance.

Furthermore, TVS-Net employs the same connectivity-preserving loss function used in Chapter 4, specifically designed to minimize segmentation discontinuities and preserve the structural integrity of the vascular tree. This loss function, crucial for achieving accurate vessel skeletonization, is further validated in this chapter by demonstrating its effectiveness in a new context where temporal information is incorporated. Additionally, the architecture integrates deep supervision, enhancing both the convergence rate and overall segmentation accuracy.

The evaluation of TVS-Net on both the JR and SJTU datasets illustrates its superior performance relative to contemporary state-of-the-art (SOTA) methodologies. This system attains a high level of segmentation quality even in challenging scenarios, such as bifurcations and areas characterized by low contrast, thereby underscoring its potential for dependable clinical applications.

Contributions

To overcome the challenges of coronary vessels segmentation in ICA, I introduce a novel TVS-Net architecture. My main contributions can be summarized in four aspects:

- The development of a new 3D (2D+T) framework that simultaneously extracts features from multiple consecutive ICA frames to segment the target frame. My framework combines a novel densely nested 3D encoder that expands through additional convolutional nodes in its mid-layers and a highly connected 2D decoder. This dual-pathway design uses UNet++ as its backbone and ensures robust spatial-temporal feature extraction and precise spatial recognition, resulting in high-fidelity segmentation.
- I further incorporate the connectivity-preserving loss function to ensure vascular structural integrity while simultaneously disentangling spatial and temporal infor-

mation. Additionally, I utilize specific skeletonized metrics to evaluate structural accuracy.

- The TVS-Net outperforms comparable approaches across varied data sources and annotation protocols, achieving 83.4% Dice and 84.3% recall on the SJTU dataset. It achieves a higher recall of 86.3% on a refined subset with fine-grained annotations, notably surpassing the original dataset's manual annotations.
- TVS-Net attains 78.5% Dice and 82.4% recall on the JR dataset comprising 60 ICAs in out-of-distribution (OOD) evaluation, outperforming all SOTA methods. These results highlight the robustness and generalizability of my approach.

5.2 Materials and Methods

This section introduces the input data, the neural network architecture, the loss function, post-processing, computational methods, and evaluation metrics. Let I_n denote the set of ICA image sequences in the dataset with a total number of N cases. Each $I_n \in [0, 255]^{T\times H\times W}$ represents a 3D sequence of T frames of height H and width W. The t-th frame of the n-th case is denoted by $I_{n,t} \in [0, 255]^{H\times W}$, and the selected frame for manual annotation is $I_{n,0}$. The corresponding gold standard segmentation is denoted by $G_n \in \{0, 1\}^{H\times W}$. The full labeled dataset is written as $D = (I_n, G_n)_{n=1}^{N}$. The segmentation function of the network is denoted by $f(\cdot)$, producing predicted segmentation maps $f(I_n)$ for each ICA sequence.

5.2.1 Study Population

For training and inference of the proposed vessels segmentation algorithm, I use the SJTU dataset D_2 from [34], acquired from the Renji Hospital of Shanghai Jiao Tong University (SJTU). This publicly available dataset contained 323 short sequences, each containing two frames before and one after each annotated frame. Hence, the dataset contains $T = 4$ frames for every annotation with $I_n = \{I_{n,-2}, I_{n,-1}, I_{n,0}, I_{n,1}\}$ centered around the annotated frame $I_{n,0}$. Hao et al. [34] experimentally demonstrated that $T = 4$ provides the best generalizability in performance of the SOTA SVS-Net compared to $T = 2, 3$, and

5. For direct comparison, I divide the 323 samples into train, validation, and test sets as $N_{train} = 173$, $N_{val} = 82$, and $N_{test} = 68$, respectively, as in [34].

I further utilize the dataset D_1 obtained from the John Radcliffe (JR) Hospital, Oxford University Hospitals NHS Foundation Trust, for out-of-distribution (OOD) experiments. In order to align the I_n of D_2, two frames preceding and one frame subsequent to each annotated frame are also extracted. All images are employed for OOD evaluation, resulting in a total of $N_{test} = 60$.

5.2.2 TVS-Net and TVS-Net+

The usage of an encoder-decoder structure and skip-connections has been proven effective in UNet [16]. To fully exploit this framework, U-Net++ [88] fills the space between the encoder and decoder with extra skip-connected nodes, generating a denser network for multi-scale structure concatenation. The 2D U-Net++ has been applied for single-frame whole vascular tree segmentation and semantic segmentation in [184]. In order to incorporate and analyze temporal information, I introduce the TVS-Net and TVS-Net+ models in this work. Since the dataset consists of multiple ICA sequences with one annotated target frame for each sequence, in order to analyze the 3D input and 2D output structures simultaneously, the network incorporates 3D and 2D convolution blocks to build a 3D encoder and a 2D decoder that are densely fused together by temporal feature extraction. A 3D convolution block contains convolution, batch normalization, and rectified linear unit (rectified linear units (ReLU)) with the kernel size for convolution being $(3, 3, 3)$ as shown in Fig. 5.1. A similar pipeline is applied in the 2D convolution block, which is the same as in Fig. 4.1 of the previous Chapter.

Specifically, the 3D encoder captures both spatial and temporal features from ICA sequences, treating time as a third dimension to enhance vessel segmentation. The 3D-2D temporal block then fuses these features across frames, enabling the 2D decoder to produce accurate vessel masks informed by the sequence's dynamics. This design helps mitigate noise from overlapping vessels, organs, and variations in contrast by using temporal information to distinguish target vessels and reduce artifacts.

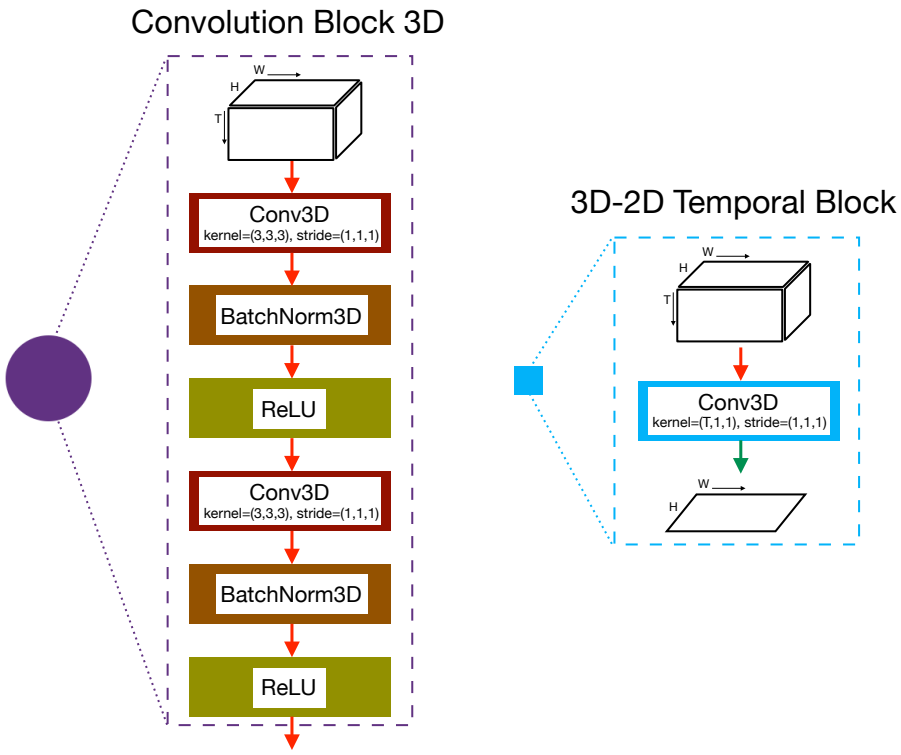


Figure 5.1: The 3D convolution block (left) and temporal feature extraction block (right). The green and red arrows represent the 3D and 2D paths separately.

Model Variants

The convolution node in the network is represented as $U^{i,j}$, with i representing the number of pooling or down-sampling operations from $U^{0,0}$ and j representing the number of received skip-connections of the node, which is the same for a 2D UNet++. In a 5-level pyramidal structure, $i, j \in \{0, 1, 2, 3, 4\}$, the number of channels in the output of each node equals 2^{i+5}. With the six interlinked nodes ($U^{0,1}, U^{0,2}, U^{0,3}, U^{1,1}, U^{1,2}, U^{2,1}$), the connection from $U^{0,0}$ to any node can be seen as a dense convolution block. For example, node $U^{1,2}$ receives two skip-connection outputs, and one up-sampling output concatenated together, where these outputs are generated in different convolution layers and pyramid levels.

The network is structured with 3D convolution blocks in the encoder and 2D convolution blocks in the decoder, which are accompanied by temporal feature extraction that always happens right after the encoder. I compare two model variants for the six interlinked nodes, as the 3D encoder can expand through these nodes.

The first variant, TVS-Net, is shown in Fig. 5.2, where those 6 nodes that belong to the dense 2D decoder can be mathematically expressed as:

$$U^{i,j} = \begin{cases} V_{3D}(U^{i-1,j}) & j = 0 \\ V_{2D}(J(F(U^{i,0}), R(F(U^{i+1,j-1})))) & j = 1 \\ V_{2D}(J(J(U^{i,k})_{k=1}^{j-1}, F(U^{i,0}), R(U^{i+1,j-1}))) & j > 1 \end{cases} \tag{5.1}$$

where functions $V_{3D}(\cdot)$, $V_{2D}(\cdot)$, $R(\cdot)$, $F(\cdot)$, and $J(\cdot)$ represent the 3D and 2D convolution blocks, up-sampling layer, temporal feature block, and concatenation layer, respectively.

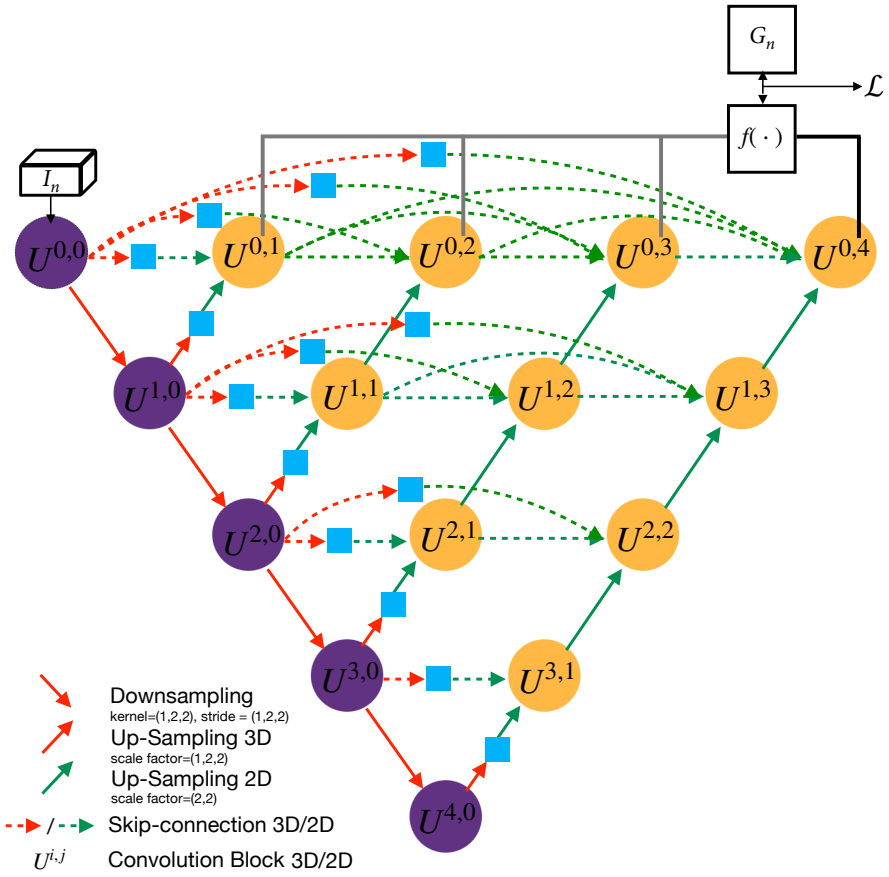


Figure 5.2: The proposed TVS-Net model for vessels segmentation from ICA sequences using temporal information. 3D and 2D blocks are represented in purple and orange, respectively, with all kernel sizes and strides shown in the boxes. The grey output paths at the top indicate deep supervision.

To elaborate, for nodes with $j = 0$, solely 3D convolution operations are performed. When $j = 1$, 2D convolution is conducted over the concatenation of the outputs of the 3D-2D temporal blocks at the same level and at the lower level. For nodes where $j > 1$, 2D convolution is executed over the concatenation of the output of the 3D-2D temporal

block at the current level, the upsampled output of the 2D node from the lower level, and all 2D nodes that are skip-connected to that node.

In the second variant, named TVS-Net+, the architecture is modified by substituting the original six nodes in the mid-layers with 3D convolutional blocks shown in Fig. 5.3, thereby creating a denser 3D encoder for spatial-temporal feature extraction. This alteration also entails rearranging the temporal feature blocks since these blocks follow the 3D convolution. Similar to Eq. (5.1), I express it as:

$$U^{i,j} = \begin{cases} V_{3D}(U^{i-1,j}) & j = 0 \\ V_{3D}(J(J(U^{i,k})_{k=0}^{j-1}, R(U^{i+1,j-1}))) & i + j < 4, j > 0 \\ V_{2D}(J(J(F(U^{i,k}))_{k=0}^{j-1}, R(U^{i+1,j-1}))) & i + j = 4, j > 0 \end{cases} \tag{5.2}$$

with the same notations. Both variants exhibit highly interconnected 3D encoder-2D decoder architectures.

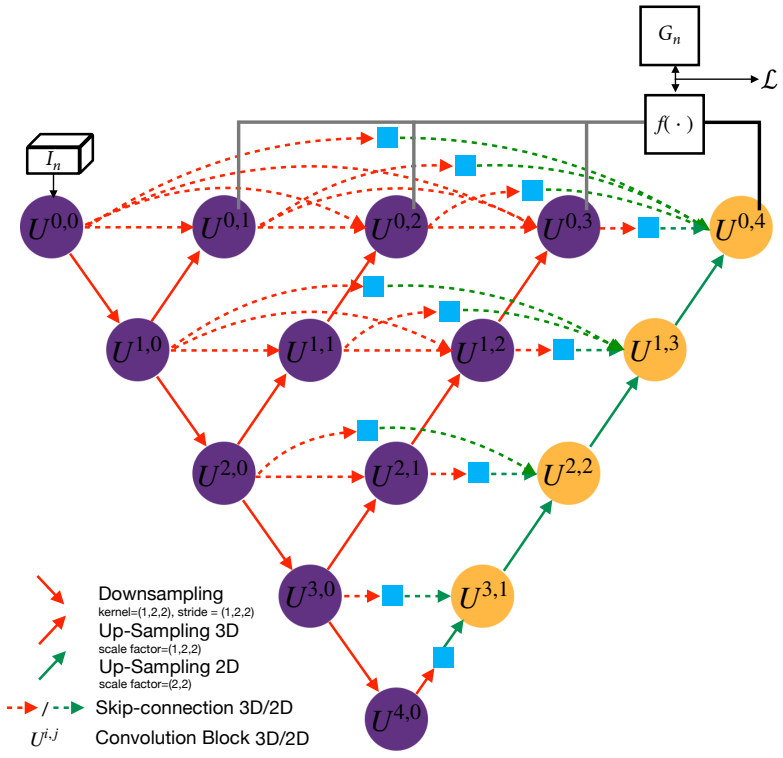


Figure 5.3: The proposed TVS-Net+ model for vessels segmentation from ICA sequences using temporal information.

However, TVS-Net features a denser 2D decoder, while TVS-Net+ emphasizes a denser 3D encoder. These configurations impart distinct biases, with the first variant opti-

mizing for spatial-temporal analysis and the second variant enhancing spatial recognition. Additionally, TVS-Net+ is larger in size compared to TVS-Net, with the extent of the size difference depending on the number of frames T in the input ICA sequence.

Temporal Feature Extraction

Time is treated as the third dimension in the model. As shown in the 3D-2D temporal block in Fig. 5.1, I first apply a 3D convolution with kernel size $(T, 1, 1)$. Then I perform dimensional compression, i.e., squeezing, on the produced feature map in its time axis, resulting in a 2D temporally fused output, allowing the model to simultaneously analyze spatial features while taking into account the temporal dynamics inherent in ICA sequences. Mathematically, this is expressed as:

$$F(\cdot) = Squeeze_T(U^{i,j} \circledast \theta_{fus}) \tag{5.3}$$

where θ_{fus} is the learnable kernel weight with size $(4, 1, 1)$ in this case and $\circledast$ denotes a convolution along the temporal dimension. This temporal extraction enhances accurate vessel segmentation by enabling consistent identification of vascular structures across frames. As a result, vessels that may appear faded or moved into low-contrast regions in the annotated frame can still be accurately detected based on their visibility in other frames without explicitly modeling vessel motion.

Deep Supervision

I incorporate deep supervision [190] in this framework, represented by the grey paths at the top of the network in Fig. 5.2 and Fig. 5.3. I first calculate the loss after applying the sigmoid function on the outputs of blocks $U^{0,1}$, $U^{0,2}$, $U^{0,3}$, and $U^{0,4}$. The final deep supervision loss is derived as the average of the four loss values, serving as an additional regularization to diminish error and expedite loss convergence.

5.2.3 Energy Loss Function

As in Chapter 4, I minimize a loss function inspired by the elastic energy of dislocations in crystals, originally proposed in [188]. For a single curve $\gamma(x(s), y(s), z(s))$ in 3D space

(x, y, z), the system energy is defined by:

$$\vec{w}(x,y,z) = -\frac{1}{4\pi}\int_{\gamma}\frac{\vec{r}\times d\vec{l}}{|\vec{r}|^3} \text{ and } E = \frac{1}{8\pi}\int_{\gamma}\int_{\gamma'}\frac{d\vec{l}\cdot d\vec{l'}}{|\vec{r}|}, \tag{5.4}$$

where $d\vec{l}$ represents a differential curve element, and $\vec{r}$ is the vector between points in space. $\vec{w}(x, y, z)$ governs system dynamics, while E represents total elastic energy. To adapt this to vessel segmentation, I define the gold standard boundary as γ_1 and the predicted boundary as γ_2. By substituting $\gamma = \gamma_1 \cup \gamma_2$ and convolving γ_1 with a Gaussian function (G_s) and γ_2 with a Heaviside function ($\eta_{step}(\phi)$), the dynamic energy minimization equation becomes:

$$v(x,y) = -\frac{1}{4\pi}\int_{\mathfrak{R}^2}\frac{\vec{r}\cdot\nabla(G_s + \alpha\eta_{step}(\phi))}{|\vec{r}|^3}dxdy, \tag{5.5}$$

where α is a hyperparameter. The corresponding energy is expressed as:

$$E = \frac{1}{8\pi}\int_{\mathfrak{R}^3}dxdy\int_{\mathfrak{R}^3}\frac{\nabla\mathbb{T}(x,y)\cdot\nabla\mathbb{T}(x',y')}{|\vec{r}|}dx'dy', \tag{5.6}$$

where $\mathbb{T}(x, y) = G_s + \alpha\eta_{step}(\phi)$. Here, G_s is the vessel boundary in the ground truth and can be obtained by convolving G_n with a 2D Gaussian function. The term $\eta_{step}(\phi)$ is a regularised Heaviside function that represents the vessel boundary in the generated segmentation. The variables x and y correspond to the column and row indices in the image grid $H \times W$. Eq. (5.6) defines the loss function $\mathcal{L}$, with its gradient expressed in Eq. (5.5). Unlike the previous chapter that applied this loss function to single-frame ICA images, this method leverages the loss function not only to capture spatio-temporal dependencies but also to preserve connectivity across consecutive frames. For computational efficiency, these equations can be further simplified in 2D Fourier space [109].

5.2.4 Half Tensor Training

Since the proposed architecture is a dense quasi-3D network, I employ half-tensor training with mixed precision to optimize segmentation performance while maintaining a smaller model size. Mixed precision training leverages the computational efficiency of 16-bit floating point (FP16) arithmetic while preserving critical accuracy in certain operations by using 32-bit floating point (FP32).

In this method, all operations except the loss calculation are performed in FP16. This includes forward and backward passes, weight updates, and intermediate tensor

computations. Using FP16 significantly reduces memory consumption and increases training speed due to smaller data sizes and faster GPU processing. However, the loss calculation is performed in FP32 to prevent numerical instability. The loss function often involves summations and products that can lead to rounding errors or precision loss when using FP16, especially for large-scale datasets or during prolonged training. Performing the loss calculation in FP32 ensures accurate gradient computations and stable optimization.

Specifically, only the gold standard segmentation masks G_n are stored in FP32 throughout the training process, as they serve as the reference for loss computation. The segmentation output $f(\cdot)$, initially generated in FP16, is converted to FP32 during the loss calculation to match the precision of the gold standard. This mixed-precision approach balances computational efficiency and numerical accuracy, enabling the model to retain high segmentation performance while reducing its overall size and training overhead.

5.2.5 Post-Processing

Skeletonization

In order to evaluate the performance of the proposed method for generating accurate vessel skeletons, the gold standard segmentation G_n and predicted segmentation $f(\cdot)$ are both skeletonized. I apply the skeletonization method $\sigma(\cdot)$ [191], which includes clockwise pixel assessment of a 3×3 mask iterated through the image. It relies on counting the number of non-zero neighbors and transitions from 0 to 1 in the surrounding pixels to determine whether a pixel should be retained. Applying this process across the binary image yields a thinned representation of the vascular structure suitable for skeleton-based evaluation.

Connected Component Filtering

In order to eliminate spurious components, I identify connected components in the predicted segmentation $f(\cdot)$ and remove those smaller than a specified threshold.

5.2.6 Evaluation Methods

Conventional Metrics

To evaluate segmentation performance, I use three conventional metrics: area under the precision-recall curve (AUPRC), Dice coefficient, and recall. This allows for direct comparison with SVS-Net [34] on the same 4-frame ICA sequence dataset. For this current study, false positives are relatively less informative than false negatives, especially since the dataset D_2 is coarsely annotated and missing smaller and thinner vessel structures. Therefore, I focus on AUPRC and recall for a comprehensive assessment. AUPRC quantifies model performance by calculating the area under the curve obtained by plotting precision against recall across varying binarization thresholds, since the model outputs a probability map rather than binary predictions. It is defined as:

$$\text{AUPRC} = \int_0^1 \text{precision}(\text{recall}), d(\text{recall}) \tag{5.7}$$

where precision evaluates how many of the pixels predicted as vessels are actually in the vessel regions of the gold standard. It is given by:

$$\text{precision} = \frac{TP}{TP + FP} \tag{5.8}$$

with TP and FP introduced in Section 4.2.6.

Skeletonization Metrics

In the previous chapter, I evaluated structural correctness using conventional metrics, such as Dice and recall, alongside supplementary metrics including over-segmentation ($\mathcal{H}(f|G)$), under-segmentation ($\mathcal{H}(G|f)$), and Betti number error. While these metrics provided valuable insights into segmentation quality and topological consistency, they possess certain limitations. Specifically, Betti number error, although effective for quantifying topological correctness, lacks an upper bound, making it challenging to interpret its values within a constrained or normalized range. This can lead to difficulties in comparing results across models and datasets.

To address these limitations, I adopt skeletonization-based metrics, completeness (C_r), and correctness (C_p), initially proposed for evaluating curvilinear structures [192]. These metrics, which can be viewed as buffered versions of recall and precision, are

well-suited for assessing the geometric preservation of vessel centerlines. Over the vessel skeleton, true/false positives/negatives are defined as illustrated in Fig. 5.4.

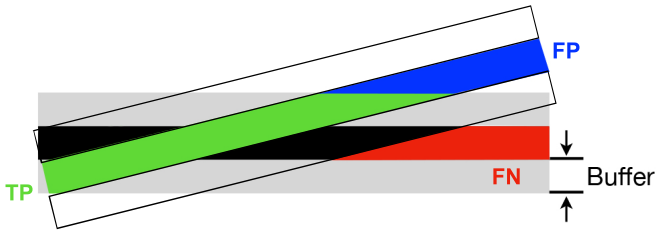


Figure 5.4: Skeletonization metrics for vessel centerline with 1 pixel width. True positive (TP) is green, whereas false positive (FP) and false negative (FN) are blue and red, respectively. The skeleton at the back is the gold standard, and the skeleton at the front is the generated skeleton.

Since the skeleton is only 1 pixel wide, I set the buffered width on both sides to 1 pixel, making the buffered skeleton 3 pixels wide. To obtain TP and FP, I buffer $\sigma(G_n)$ only to compare with the unbuffered $\sigma(f(\cdot))$. The part within the buffered $\sigma(G_n)$ is considered as TP and vice versa for FP. Equivalently, I obtain FN by solely buffering $\sigma(f(\cdot))$. The buffering of skeletons is performed by dilation of the single-pixel structure.

Switching to these metrics enables us to focus on the accuracy of vessel skeletonization, which directly relates to the structural preservation required for downstream applications like 3D reconstruction. Moreover, by constraining the evaluation space to the skeleton and its buffered region, C_r and C_p provide interpretable and bounded values, offering a more standardized and robust assessment.

5.3 Experimental Results

In this section, I introduce the experimental settings for training the proposed network and present extensive quantitative and qualitative comparisons. Finally, I conduct a comprehensive study on the re-segmented gold standard.

5.3.1 Experimental Settings

The training of the network is performed on an NVIDIA Quadro RTX 8000 GPU, with all frames augmented on-the-fly by flipping, changing saturation, and changing contrast. I also apply rotation by 90° randomly for 0-3 times with probability 0.7. Trainings in full tensor

are optimized with Adam using a learning rate of 10^{-5}, $\beta_1 = 0.85$, weight decay of 10^{-5}, and all other hyperparameters set to the default in Pytorch. Trainings in half tensor are optimized with Stochastic Gradient Descent (SGD) to avoid the gradient vanishing problem in Adam optimization when values exceed the dynamic range of FP16. To support SGD, I apply a cosine annealing scheduler to gradually decrease the learning rate from 10^{-6} to 10^{-7}. Besides, the weight decay is decreased to 10^{-6} and momentum is set to 0.9 for optimization with SGD. For all training, the value of α in the loss function of Eq. (5.6) is 0.35. Hyperparameters are obtained by grid search. I ran training for over 2000 epochs to ensure convergence of the loss function.

5.3.2 TVS-Net and TVS-Net+

To compare the effectiveness of temporal information for multi-frame ICA segmentation, I train with the same hyperparameters in two proposed network architectures: TVS-Net and TVS-Net+. This experiment is separated from other experiments since, due to the large network size for TVS-Net+, the batch size for this experiment is limited to 4. As presented in Table 5.1, the TVS-Net produces the best results with the highest AUPRC and Dice along with lower standard deviation (SD).

Table 5.1: Comparison of different architecture variants.

Model	AUPRC	Dice (%)	Recall (%)
TVS-Net	**0.8717**	**82.65$\pm$2.56**	82.20$\pm$4.68
TVS-Net+	0.8653	82.28$\pm$2.57	**82.85$\pm$5.73**

All values represent mean $\pm$ standard deviation (SD).

The performance of these two networks can be visually assessed in Fig. 5.5. From the red arrows in the first row, it can be seen that the TVS-Net not only yields a good delineation of thick vessels but also accurately defines small bifurcations and the distal part of the coronary vascular tree, pivotal for preserving the vascular topology. This, along with the computational efficiency and performance improvement, suggests the relative superiority of the proposed TVS-Net model in this study.

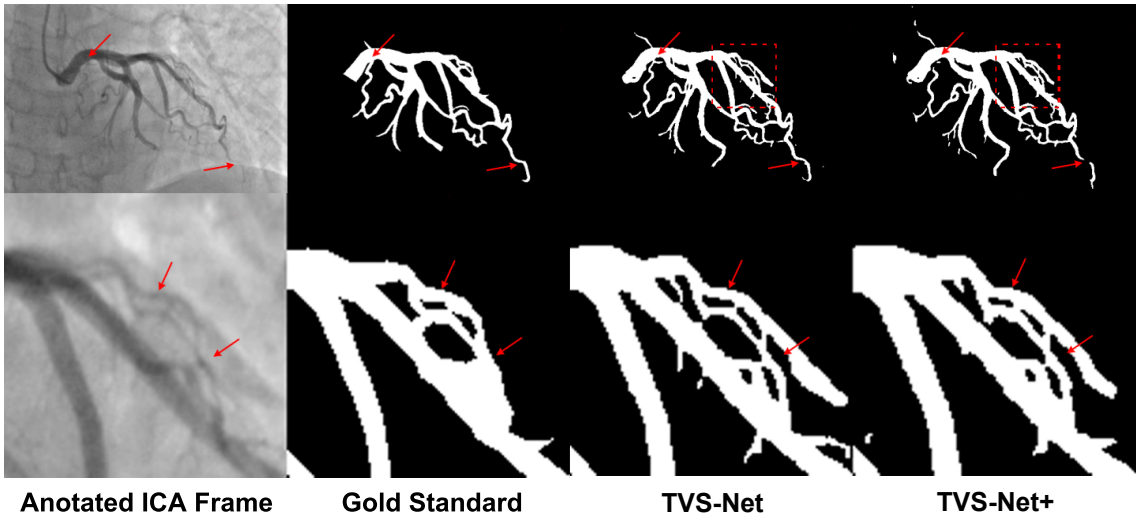


Figure 5.5: Qualitative evaluation of segmentation performance using TVS-Net and TVS-Net+. The bottom row is a zoomed-in version of the red square in the top row. The dotted boxes are zoomed regions. The arrows highlight where the prediction surpasses the gold standard in coarse-grained regions.

5.3.3 Comparison Methods

To evaluate the performance of the proposed TVS-Net framework, I compare 6 SOTA models that have been applied in vessel segmentation, namely UNet [16], RA-UNet [193], UNet++ [88], nnUNet [90], VMUNet [96], and SVS-Net [34] on the SJTU dataset D_2 and the JR dataset D_1, with SVS-Net being the SOTA temporal ICA segmentation method. Specifically, SVS-Net employs the same temporal input data but does not utilize deep supervision, whereas the other methods neither use temporal input data nor deep supervision. The loss functions used for training are also consistent with their original implementations: UNet, UNet++, and SVS-Net are trained using Dice loss, while nnUNet and VMUNet utilize a combination of Dice and cross-entropy losses. Since removing the 3D-2D temporal blocks effectively transforms TVS-Net to a UNet++, this makes UNet++ a suitable comparison model for evaluating the efficacy of temporal feature extraction in TVS-Net. For out-of-distribution (OOD) evaluation on the D_1 dataset, I directly apply the model trained on D_2. All quantitative metrics that require binarization of the generated segmentation apply the threshold of 0.5 (equivalent to 127 as pixel intensity). The optimal batch size for TVS-Net is determined to be 6, which is followed for all subsequent experiments.

5.3.4 Evaluation on the SJTU Dataset D_2

Figure 5.6 presents the segmentation performance on the test set of D_2, showcasing complex views of the left circumflex artery (LCX) and right coronary artery (RCA). It is evident that TVS-Net surpasses other SOTA methods by effectively preserving vascular structures in the main and distal branches. In the particularly narrow vessel areas of the first two images for LCX, it identifies vessels absent from the gold standard annotations, accurately capturing the width of each branch. Moreover, the performance disparity in the third image for RCA is notable, primarily because dataset D_2 comprises only about 25% of RCA images. This limitation restricts weaker frameworks from effectively transferring learned features from other images. Extreme samples can be observed in the output of UNet and RA-UNet as they fail to generalize on RCA with catastrophic disconnection and the absence of major branches.

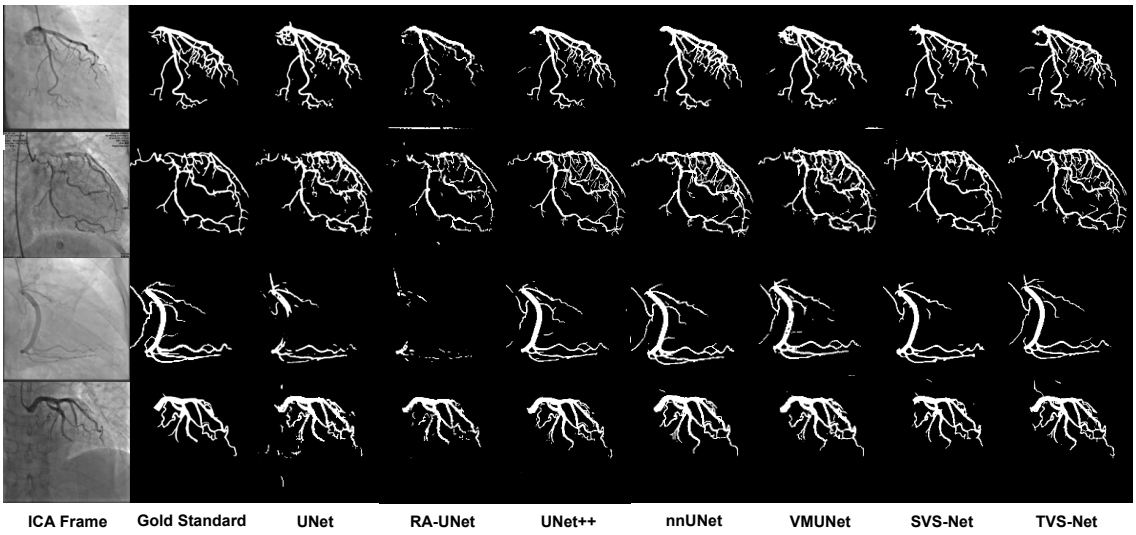


Figure 5.6: Qualitative evaluation of segmentation performance of SOTA methods and TVS-Net on dataset D_2.

Quantitatively, as shown in Table 5.2, TVS-Net achieves a Dice score of 83.41%, compared to 81.63% by VMUNet and 83.23% by SVS-Net. The improvement in Dice over VMUNet was statistically significant (p-value < 0.001). Moreover, TVS-Net demonstrated an improvement in recall to 84.31%, a 6.2% increase over SVS-Net, showcasing its ability to detect thin vessel branches that other methods often fail to identify.

In terms of skeletonization metrics, TVS-Net maintained remarkable consistency between recall and C_r, with a minimal drop from 84.31% to 84.83%, indicating its ability

to preserve skeleton integrity without significant disconnections. In contrast, methods such as nnU-Net showed a marked decline in skeletonization metrics (recall: 83.41% to C_r : 79.72%), demonstrating their inability to maintain vessel connectivity.

Table 5.2: Comparison of TVS-Net with SOTA methods on test dataset of D_2.

Method	AUPRC	Dice (%)	Recall (%)	$C_r(\%)$	$C_p(\%)$
UNet	0.8169	79.34±4.07***	79.31±7.99***	76.08±8.59***	71.96±8.05***
RA-UNet	0.8109	72.13±9.51***	62.22±12.0***	61.57±12.2***	*78.47±6.34*
UNet++	0.8687	81.54±3.57***	81.13±6.59***	80.99±8.49***	75.02±7.19**
nnUNet	0.8679	81.64±4.05**	83.41±8.74	79.72±9.17*	73.32±8.51*
VMUNet	*0.8839*	81.63±2.85***	*83.68±4.53**	*83.03±5.27****	74.72±5.69**
SVS-Net	0.8621	*83.23±4.49*	79.41±7.46***	77.32±7.54***	**82.71±6.97**
TVS-Net	**0.8899**	**83.41±2.45**	**84.31±4.62**	**84.83±5.74**	76.55±7.29
TVS-Net w Dice w/o DS	0.8826	82.49±2.81*	84.60±4.45	81.92±5.04**	75.63±7.02
TVS-Net w/o DS	0.8847	83.27±2.57	83.49±4.41*	83.52±5.54	76.87±7.03

All values represent mean ± SD. * p-value < 0.05, ** p-value < 0.01, and *** p-value < 0.001, based on one-tailed Wilcoxon signed-rank test against the TVS-Net.

5.3.5 Evaluation on the JR Dataset D_1

For additional quantitative evaluation of the generalizability of the proposed TVS-Net, I apply the trained model from the dataset D_2 on the out-of-distribution (OOD) dataset D_1 acquired from the Oxford JR Hospital. As visible from the results presented in Table 5.3, TVS-Net achieves superior performance in terms of all metrics AUPRC, Dice, recall, C_r, and C_p.

Table 5.3: Comparison of TVS-Net with SOTA methods on OOD dataset D_1.

Method	AUPRC	Dice (%)	Recall (%)	$C_r(\%)$	$C_p(\%)$
UNet	0.7207	67.88±18.1***	67.86±21.6***	58.88±20.1***	68.49±15.2***
RA-UNet	0.6746	51.18±18.5***	39.12±18.0***	39.43±17.9***	73.59±15.3*
UNet++	0.8030	*75.71±10.4****	75.65±13.3***	78.62±14.7***	69.52±12.4***
nnUNet	*0.8189*	74.81±8.99***	79.03±9.56***	81.81±15.1***	69.23±12.7***
VMUNet	0.8109	74.69±9.07***	*80.74±12.2**	*83.90±13.1****	*73.68±11.1****
SVS-Net	0.7642	73.64±9.72***	76.44±11.9***	71.79±12.5***	67.59±13.4***
TVS-Net	**0.8437**	**78.49±9.74**	**82.41±11.8**	**86.53±13.5**	**77.78±11.7**

All values represent mean ± SD. * p-value < 0.05 and *** p-value < 0.001, based on one-tailed Wilcoxon signed-rank test against the TVS-Net.

The Dice score achieved by TVS-Net is significantly higher than the next-best method (VMUNet, p-value< 0.001). Similarly, the recall of TVS-Net surpasses that of VMUNet

by 2%, with a statistical significance of p-value< 0.05. The skeletonization metrics C_r and C_p further illustrate the robustness of TVS-Net in maintaining vascular structural integrity, with C_r showing a 4.6% improvement and C_p a 7.0% improvement compared to the second-best results. These p-values demonstrate that the improvements in metrics such as Dice, recall, and C_r are statistically significant, reinforcing the efficacy of TVS-Net over competing frameworks. This statistical significance highlights the consistency and reliability of TVS-Net's performance across diverse datasets, especially in the OOD context.

In Fig. 5.7, the precision-recall (PR) curve of TVS-Net lies above all other PR curves, demonstrating a dominant generalizability for all binarization thresholds. Qualitatively, a similar observation is evident in Fig. 5.8. All other frameworks produce noisy or incomplete vascular structures in the first three RCA images. Most importantly, even SOTA SVS-Net and VMUNet cannot delineate the distal branches in the first and third images. While SOTA nnUNet manages to detect some of the finer vessels, it struggles with false positives, mistakenly identifying the catheter as a vessel in the third image. Hence, the high-performance improvement and accurate segmentation quality together demonstrate the TVS-Net's ability to generalize in OOD real-world scenarios.

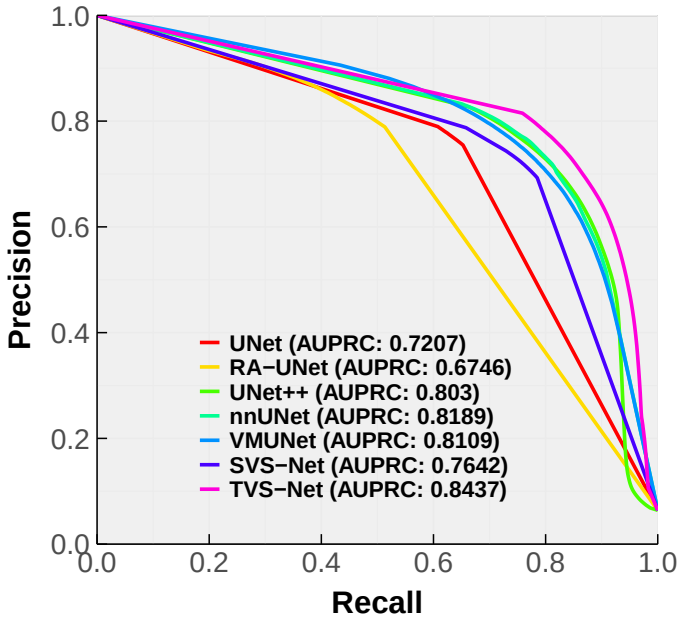


Figure 5.7: Precision-recall curve for evaluation on OOD dataset D_1.

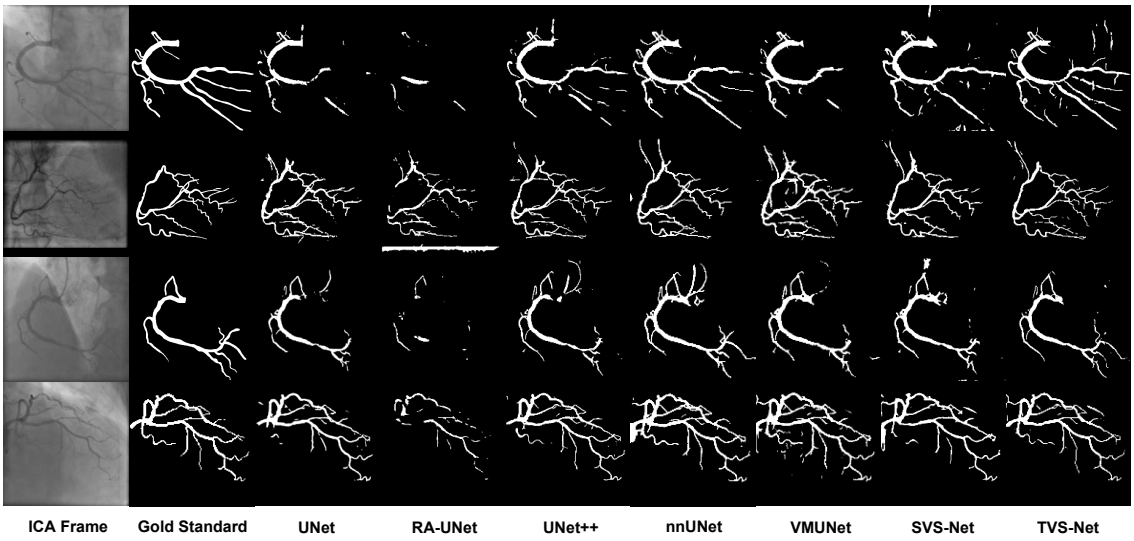


Figure 5.8: Qualitative evaluation of segmentation performance of SOTA methods and TVS-Net on OOD dataset D_1.

5.3.6 Efficacy of Energy Loss function

To evaluate the performance gain from the energy loss function $\mathcal{L}$ in Eq. (5.6), I train TVS-Net with Dice loss and the energy loss separately, without deep supervision in both cases. This approach allows for a clearer evaluation of the individual contribution of each component. From the last two rows of Table 5.2 and Fig. 5.9, I find that the energy loss function improves Dice by 1%. The recall with Dice loss is slightly higher, which may be caused by over-segmentation of the vessel boundaries, e.g. in the zoomed area of the first row in Fig. 5.9 near the arrows.

Additionally, the Dice loss causes disconnections on the main branch of the first image in Fig. 5.9, which not only affect the geometry but also yield 1.95% lower C_r and 1.64% lower C_p for skeletonized vessels (Table 5.2). Furthermore, I observe an increase from recall to C_r when applying the energy loss, indicating that the vessel skeleton and boundaries are more consistently preserved with this loss function. Notably, the elastic loss better preserves the vascular geometry even in false positives (which could potentially be true positives if a more comprehensive manual segmentation was performed), as highlighted in the boxed area of the second row in Fig. 5.9.

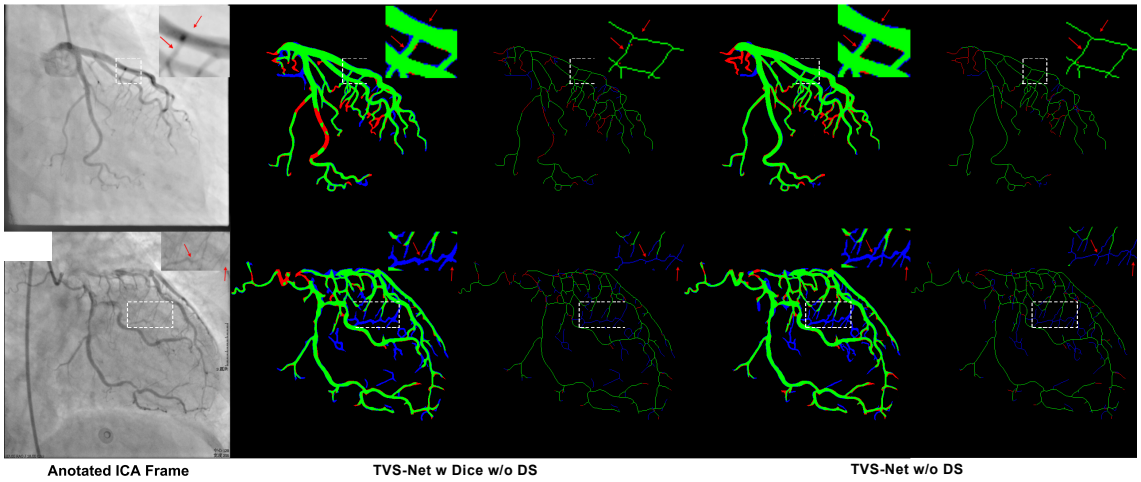


Figure 5.9: Qualitative evaluation of segmentation and skeletonization performance of TVS-Net with Dice loss and energy loss, without deep supervision. The color codes for TP, FP, and FN are the same as in Fig. 5.4. The red arrows highlight the over-segmented vessels in the zoomed region.

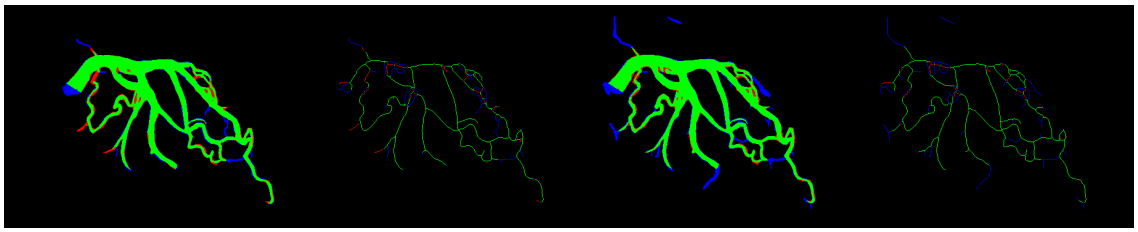
5.3.7 Effectiveness of Deep supervision

The advantage of deep supervision (DS), introduced in Section 5.2.2, is presented in Table 5.2, comparing TVS-Net to TVS-Net w/o DS. With the use of deep supervision, both Dice and recall metrics increase. A more pronounced trend is found for skeletonized vessels, where the TVS-Net with DS yields higher C_r, showing that DS significantly aids in preserving vessel connectivity.

For qualitative evaluation of the effects of deep supervision, I observe in Fig. 5.10 that for the segmentation of both vascular trees and skeletons, the deep supervision has enabled the identification of previously missed vessels (red FN in the left panel) with accurate delineation (green TP in the right panel). Hence, the utilization of deep supervision enhances the detection of missed vessels and rectifies vessel disconnection, with FP occurring only at the distal end of some vessels.

5.3.8 Comparison on Fine-Detailed Segmentation

Performing a thorough manual annotation of all vessels in an ICA image is very demanding; thus, gold standard annotations are typically coarse-grained (incomplete). The zoomed-in region in Fig. 5.5 highlights this limitation, e.g., at the location indicated by the lower red arrow. This affects the reliability of evaluation metrics. Hence, I select a subset of 10



TVS-Net w/o DS TVS-Net

Figure 5.10: Vessel segmentation and skeletonization performance of TVS-Net without and with deep supervision. Color codes are the same.

samples from the test dataset and perform a more comprehensive manual re-segmentation by an expert with the tool mentioned in Chapter 3, aiming at accurately delineating all the vessels in the images regardless of their sizes. To prevent bias in the selection, I rank the test set images based on the Dice scores of their segmentation using TVS-Net, and select those at the $0, 10, 20, 30, 40, 50, 60, 75, 85, 100^{th}$ percentiles. The 10 samples are denoted by the orange bars in Fig. 5.11, with their Dice values shown on the top of each bar.

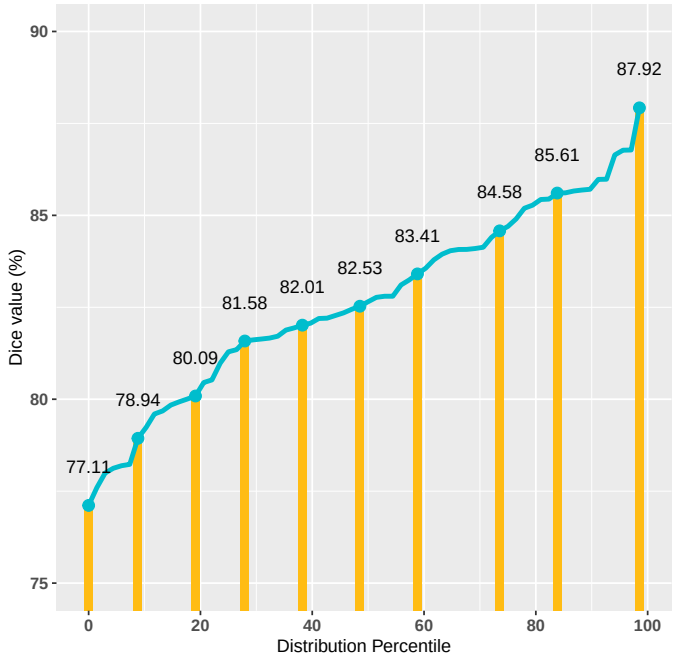


Figure 5.11: Cumulative distribution of the Dice scores by TVS-Net for all 68 test samples.

I consider this newly segmented dataset as the new “fine-grained” gold standard and compare it with the “coarse-grained” gold standard. Three example cases with the minimum (0^{th}), median (50^{th}), and maximum (100^{th}) Dice scores are shown in Fig. 5.12. From the results, it is clear that the level of incompleteness in the initial annotations

varies significantly between samples, which has an important effect on the calculated metrics. Using these 10 re-segmented samples as the new gold standard, I perform a

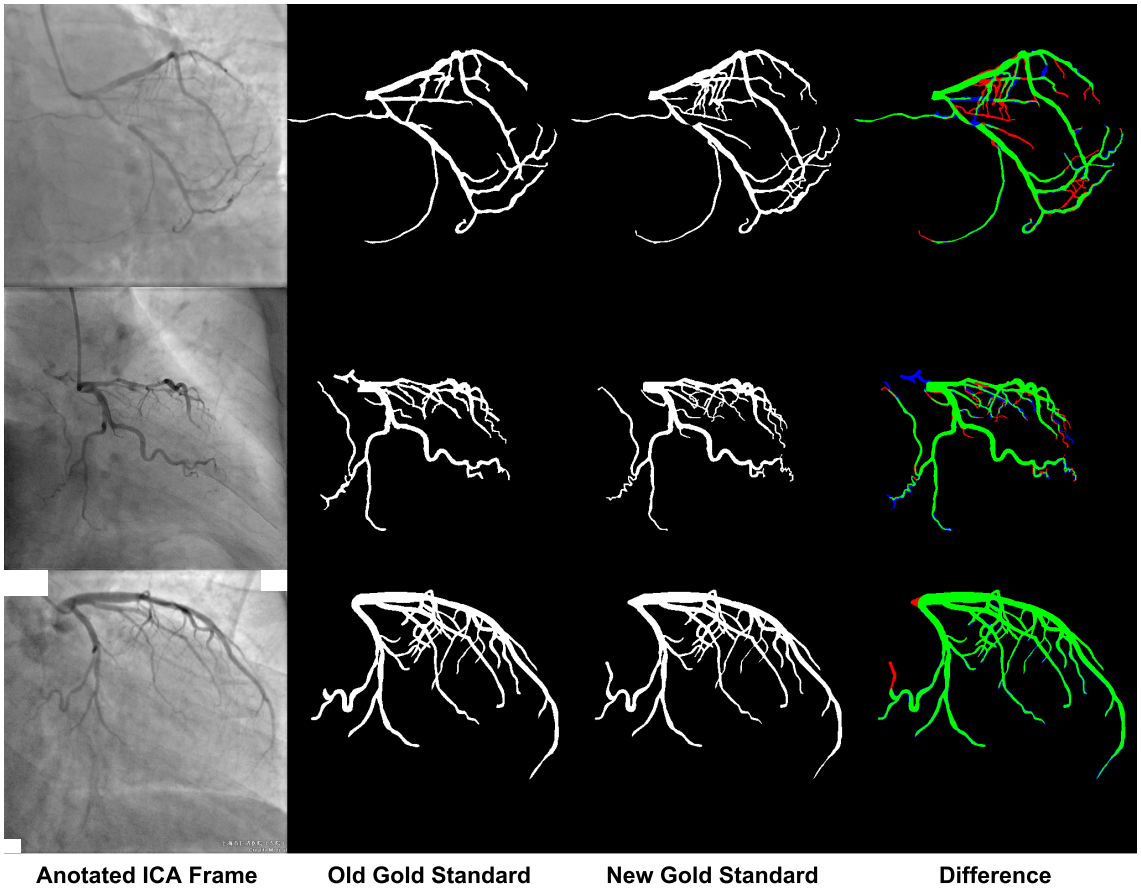


Figure 5.12: Three re-segmented samples with minimum (81.61%), median (84.69%), and maximum (94.76%) Dice scores (top to bottom). Color codes are the same.

direct evaluation using the TVS-Net trained on the original gold standard. As presented in Table 5.4, TVS-Net demonstrates significant improvements over the state-of-the-art (SOTA) temporal segmentation method, SVS-Net, achieving a Dice score increase of 7.95% and a recall improvement of 16%. Notably, the method's recall surpasses even the gold standard on which it was trained, highlighting the ability of TVS-Net to generalize and capture finer details missed in the coarse annotations of the original gold standard. The Dice score difference, however, remains a modest 1.5%, reflecting consistent segmentation quality. My method also achieves the highest completeness metric (C_r) at 84.0% while maintaining a comparable correctness metric (C_p) of 87.55%, showcasing its ability to produce more continuous and accurately aligned vessel skeletons with minimal disconnections and shifts. The p-values in Table 5.4 further substantiate these improvements, indicating

Table 5.4: Performance evaluation on the new gold standard with 10 re-segmented samples.

Model	AUPRC	Dice (%)	Recall (%)	$C_r(\%)$	$C_p(\%)$
UNet	0.8307	82.11±2.08**	81.90±5.82**	73.07±7.99**	78.28±6.75**
RA-UNet	0.8132	72.10±9.33**	61.47±11.0**	56.15±10.9**	84.91±4.40
UNet++	0.8839	*84.65±3.19**	83.07±5.42*	*82.91±5.53*	85.13±4.65
nnUNet	*0.8889*	83.42±2.41**	84.25±4.61	82.84±3.85	85.68±4.31
VMUNet	0.8877	82.12±1.92**	*85.36±4.32*	82.87±4.56	*85.93±4.83*
SVS-Net	0.8338	79.84±4.25*	74.38±8.49**	66.38±7.89**	85.74±4.39
TVS-Net	**0.9018**	**86.20±1.96**	**86.26±5.24**	**84.00±5.25**	**87.55±6.17**
Old gold standard	N/A	87.66±5.30	85.44±7.95	76.81±11.24	91.35±5.08

All values represent mean ± SD. * p-value < 0.05 and ** p-value < 0.01, based on one-tailed Wilcoxon signed-rank test against the TVS-Net.

statistical significance for all major metrics. However, the p-value is not smaller than 0.001, which is primarily due to the limited sample size of 10 re-segmented images. A larger sample size would provide more robust statistical power to detect smaller p-values. The qualitative trends are consistent with these quantitative findings, as illustrated in Fig. 5.13. Additionally, TVS-Net outperforms the SOTA single-frame segmentation method, VMUNet, with a 4.97% improvement in Dice, underscoring the efficacy of incorporating temporal information for more accurate segmentation results.

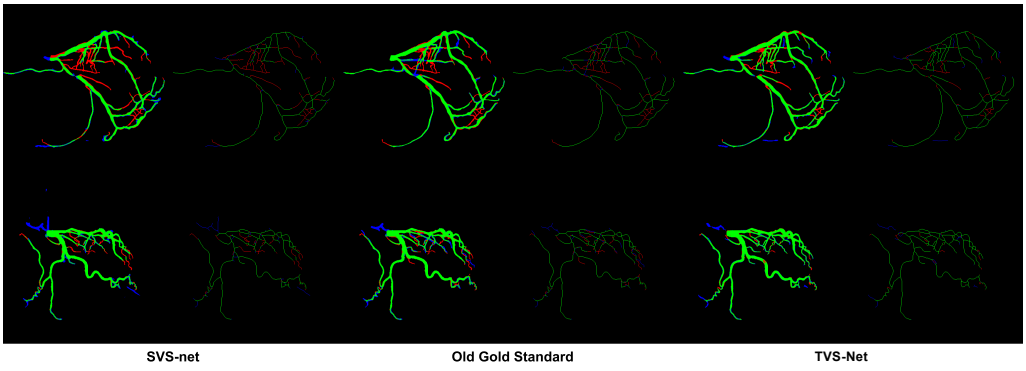


Figure 5.13: Qualitative evaluation of segmentation and skeletonization on the new gold standard. The same color code is used here.

5.4 Discussion and Conclusion

In this chapter, I develop a novel deep learning framework with a densely connected 3D encoder-2D decoder, named TVS-Net, that utilizes multiple frames of ICA sequences

for generating accurate coronary vessels segmentation. The architecture integrates temporal convolution blocks for fusing the image sequence information and a unique energy loss for enhancing topology preservation onto a dense framework featuring deep supervision. While not implemented in this study, this framework could be integrated with post-processing techniques such as the one proposed by Xian et al. [194] to further improve segmentation quality. To enable direct comparison between this method and the closest SOTA temporal segmentation [34] and other SOTA methods, I train this framework on the same SJTU dataset using the same data split. The experimental analysis not only demonstrates the advantages of a multi-frame approach but also illustrates the problem of incomplete annotations of vascular trees. The significant performance improvements observed by the TVS-Net over UNet++ underscore the effectiveness of integrating densely connected 3D-2D temporal blocks. I propose the use of skeletonization metrics and demonstrate the ability of my method to preserve thin vascular structures accurately. The statistical analyses indicate that the differences in performance metrics between TVS-Net and baseline methods are highly significant, with several metrics yielding p-values far smaller than 0.001. This reinforces the robustness of the proposed framework and the reliability of its improvements over prior methods. I propose the use of skeletonization metrics and demonstrate the ability of my method to preserve thin vascular structures accurately. The combined use of spatio-temporal encoding, connectivity-preserving loss function, and deep supervision enables the proposed TVS-Net to capture a wide range of vessel appearances and motion patterns. This generalizability is further highlighted in the evaluation using data collected from a local hospital, where my method achieves the highest Dice, recall, and AUPRC of 78.49%, 82.41%, and 0.8437, respectively, despite the fact that the hyperparameters were optimized for the SJTU dataset alone and no retraining was performed. The smaller size of the JR dataset and the random selection of cases may contribute to greater variance and performance variability. Due to the OOD nature of the JR data, all models show less confident predictions. However, the fact that my method excels in this context demonstrates its ability to adapt to different annotation protocols and data characteristics and highlights its strong potential for real-world deployment. Additionally, I proposed a variant of TVS-Net, named TVS-Net+, with the same approach after expanding the 3D encoder. However, due to the larger size of TVS-Net and available GPU resources,

its batch size is limited to 4, which impacts the network's ability to generalize effectively, thereby limiting the observed performance gains of TVS-Net+.

I re-segment 10 cases with a strict annotation protocol, including all visible vessels, and use them as the new (fine-grained) gold standard for the same evaluation pipeline. Trained by the original (coarse-grained) gold standard, my proposed TVS-Net achieves a recall of 86.26%, which is 0.96% higher than the original gold standard and 16% higher than the current SOTA method, SVS-Net [34]. The performance in accurately preserving vascular skeletons achieves 84.00% in C_r, improving on the original gold standard by 9.36% and SVS-Net by 26.54%. The qualitative evaluation illustrates the improvements, including the reduction in over-segmented vascular boundaries. The p-values from the experiments further substantiate these findings. While the limited sample size of 10 re-segmented cases restricts statistical power, TVS-Net's improvements in recall, Dice, and skeletonization metrics remain statistically significant. These results underscore the robustness of this framework, even under constrained evaluation conditions. Most interestingly, my extensive analyses demonstrate the feasibility of weak supervision with coarse-grained annotations for coronary vessels segmentation. This is evidenced by the superior delineation achieved by the TVS-Net model compared to its training gold standard, namely the coarse-grained annotations, when evaluated against fine-grained annotations. It is important to note that the ICA datasets were created by selecting high-quality frames, limiting the ability to fully demonstrate the network's performance across low-quality frames, where manual segmentation is more challenging and often results in coarser gold standards. Consequently, by modulating the completeness level of manual annotations in the dataset, this framework can also facilitate the exploration of a time-performance trade-off between manual and automatic segmentations in annotation protocols, as well as elucidate the impact of partially segmented ground truth on final trained segmentation quality.

Chapter 6

Semantic ICA Segmentation

Chapter contents

Part of this chapter was presented in the paper "Automated coronary vessels segmen-

tation in X-ray angiography using graph attention network," in International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 209–219, Springer, 2023.

6.1 Introduction

In Chapters 4 and 5, I introduced solutions for binary segmentation of invasive coronary angiography (ICA) in scenarios with limited data, and explored methods to enhance accuracy while preserving vascular structure integrity when ample data is available. However, in a clinical setting, clinicians need to focus on the vessel containing stenosis, identify the affected branch along with the branches connected to or supplying it, and isolate them from the rest of the vascular tree to better examine pathological areas. Thus, it would be highly beneficial if the segmented vessels could be divided into branches with clinical relevance, allowing clinicians to selectively view specific branches. This approach essentially constitutes the semantic segmentation of vessels in ICA.

Current research on coronary vessels segmentation from ICA has primarily focused on pixel-wise manipulation using convolutional neural networks (CNNs) [34, 110, 154]. However, CNNs learn local and spatial features while assigning pixel-wise labels on a regular grid, which may not be an optimal approach when extracting and analyzing branching structures such as vascular trees. In contrast, graph neural networks (GNNs) are specifically designed to analyze structural information represented as graphs with nodes and edges. If the vascular tree's structural information can be transformed into a graph, GNNs could be effectively utilized. Moreover, since the creation of a well-annotated binary segmentation dataset is already highly labor-intensive and time-consuming, no publicly available semantically segmented ICA dataset exists.

To this end, I first design a user interface for efficiently annotating semantic segmentations based on an available binary segmentation dataset. I build a new semantic segmentation dataset in order to address the problem of dataset scarcity. Then, I propose two approaches that effectively integrate CNNs and GNNs, achieving semantic vessels segmentation by combining local and spatial features with global and structural information, compared to a conventional CNN-based method. In the first method, I propose a novel parallel fusion architecture integrating CNN and GNN, which utilizes only the original ICA as

input. This framework leverages CNN feature maps to construct the graph and effectively combines spatial information from the CNN with structural information from the GNN, thus enhancing the semantic segmentation of vessels in ICA. The second method employs a cascaded architecture in which the GNN follows the CNN through three sequential steps: first, the CNN utilizes spatial information to perform binary segmentation; second, a binary skeleton is generated and transformed into a graph structure for input into the GNN; finally, node classification labels from the GNN are propagated back to refine the binary segmentation. In the second step, I introduce a novel graph construction method based on partitioned binary skeletons, allowing the generation of a tree-structured graph that captures hierarchical structural information of vessel branches and enables seamless propagation of node labels back onto vessel skeletons. Additionally, I propose a bespoke branching penalty loss designed specifically to enforce the characteristic absence of side branches in the main coronary vessel. In the third step, I incorporate a skeleton correction algorithm to further reinforce these structural constraints, achieving more accurate semantic segmentation.

The evaluation of the cascaded approach on the semantic segmentation dataset shows superior performance compared to a conventional CNN. Although the fusion approach only outperforms the conventional CNN in binary segmentation tasks, the cascaded approach delivers higher-quality semantic segmentation with enhanced structural integrity across classes, surpassing state-of-the-art (SOTA) semantic segmentation methods.

Contributions

My main contributions can be summarized in four aspects:

- I develop an efficient semantic annotation tool that leverages existing binary segmentation datasets, significantly reducing the effort required to generate high-quality semantic annotations for coronary vessels in ICA images.
- I propose two novel frameworks integrating CNNs and GNNs for the semantic segmentation of coronary vessels. The first is a fusion approach that dynamically merges spatial and structural features, while the second is a cascaded approach that sequentially employs CNN-generated binary segmentations and structural features derived

from skeletonization for precise node classification and semantic segmentation.

- I introduce an effective skeleton partitioning and graph-generation method that converts segmented vessel structures into hierarchical tree-structured graphs. This approach captures vessel branch relationships effectively. Additionally, I propose specialized penalty loss functions designed to improve accuracy by penalizing misclassifications at vessel branching points and minimizing error propagation along vessel branches.
- My fusion approach achieves superior performance in binary segmentation with a Dice score of 80.5% and recall of 82.0%, outperforming conventional CNN-based methods. The cascaded approach demonstrates excellent performance in the semantic segmentation of main coronary branches, reaching a Dice score of 76.1%, recall of 81.8%, and precision of 73.9%, surpassing the SOTA method.

6.2 Materials and Methods

This section introduces the input data, describes in detail the fusion and cascaded approaches, and explains post-processing, computational methods, and evaluation metrics. For each case n, I denote I_n as the ICA image sequence and G_n as the gold standard semantic segmentation, where $I_n \in [0, 255]^{T \times H \times W}$ and $G_n \in [0, 255]^{H \times W}$, with T, H, and W representing the number of temporal frames and the height and width of each frame, respectively. The segmentation label G_n is a class index label map that can be visualized as a color-coded mask, where each class corresponds to one of C unique RGB colors.

I construct a corresponding graph for each case as $\mathcal{Q}_n = (\mathcal{V}_n, \mathcal{E}_n)$, where $\mathcal{V}_n$ is the set of vessel-related nodes and $\mathcal{E}_n$ denotes their connectivity. Each node $\mathbf{v} \in \mathcal{V}_n$ is represented as a feature vector $\mathbf{v} \in \mathbb{R}^d$, where d is the number of node features. Each node is also assigned a class label $a_{\mathbf{v}} \in \{1, 2, \dots, C\}$, derived from G_n based on the node's spatial location. The set of all node labels for graph $\mathcal{Q}_n$ is denoted as $\mathcal{A}_n = \{a_{\mathbf{v}} \mid \mathbf{v} \in \mathcal{V}_n\}$.

The complete dataset is thus defined as $D = \{(I_n, G_n, \mathcal{Q}_n, \mathcal{A}_n)\}_{n=1}^{N}$, where N is the total number of cases. The pixel-wise output of the CNN is denoted as $f_{CNN}(\cdot) \in [0, 255]^{H \times W}$ and can be visualized as a color-coded segmentation mask. Likewise, the GNN predicts node-level class assignments as $f_{GNN}(\cdot) = \{\hat{a}_{\mathbf{v}} \mid \mathbf{v} \in \mathcal{V}_n, \hat{a}_{\mathbf{v}} \in \{1, 2, \dots, C\}\}$,

where $\hat{a}_{\mathbf{v}}$ denotes the predicted label for node **v**. These outputs, $f_{CNN}(\cdot)$ and $f_{GNN}(\cdot)$, are jointly leveraged to refine the final semantic segmentation by integrating both spatial and structural information from the CNN and GNN predictions.

6.2.1 Dataset Creation

Due to the complex, dynamic vascular structures, creating the gold standard G_n for coronary vessel semantic segmentation in ICA is highly labor-intensive and time-consuming. The frequent presence of overlapping vessels, a common occurrence in ICA imaging, further complicates the annotation process. Accurately assigning semantic labels in such cases requires annotators to manually review the entire image sequence to disambiguate overlapping structures, further increasing the complexity and effort required for annotation. As a result, the availability of semantically labeled ICA datasets remains highly limited, with no publicly available datasets currently existing for this task. To address this challenge and establish a gold standard for model training, rather than manually constructing a new semantic ICA dataset from scratch, I developed a semantic annotation tool (Fig. 6.1). This tool enables the efficiency of semantic labels to pre-existing binary-segmented ICA images, thereby reducing manual effort. The interface leverages vessel skeletons extracted from binary segmentation to streamline the annotation process. The datasets introduced in Chapter 3, originally binary segmentation datasets, serve different purposes. The JR dataset D_1 is utilized for semi-supervised binary segmentation, while the SJTU dataset D_2 is applied for supervised temporal binary segmentation. Both datasets can be seamlessly integrated into the annotation tool, allowing for the automated generation of semantic labels from binary segmentation, significantly reducing annotation workload while ensuring high-quality semantic annotations.

The annotation process begins by skeletonizing the binary segmentation and loading it alongside the binary segmentation mask and the corresponding ICA image. The user then initiates annotation by selecting point A (the starting point) and point B (the endpoint) on the vessel skeleton of interest using the designated buttons in Fig. 6.1. Once selected, the ”Line Gen” button is clicked, triggering the A pathfinding algorithm, which automatically highlights the vessel skeleton between points A and B [195]. A predefined color is assigned to the path according to the selected semantic vessel label. Subsequently, the ”Vessel

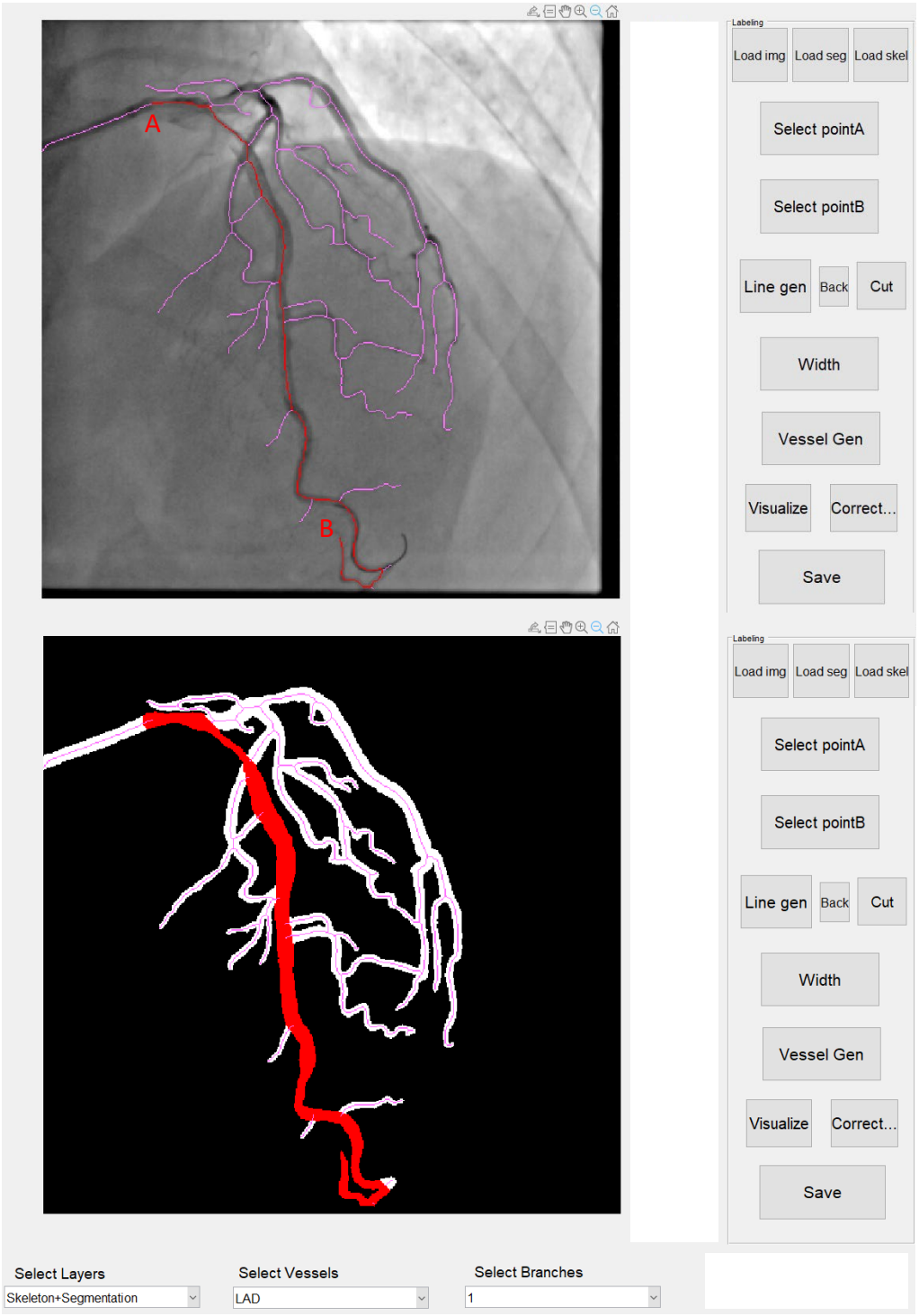


Figure 6.1: Matlab interface for semantic annotation. The pink lines represent unlabeled skeletons. The interface displays the process of labeling the LAD skeleton and its corresponding vessel.

Gen" button is used to propagate the color throughout the vascular area. This is achieved by expanding a disk centered on each skeleton point, which continues expanding until its

boundary intersects the binary segmentation mask for the second time, ensuring complete coverage of the vessel region. The combined effect of these expanding disks, constrained by the binary segmentation mask, results in the final semantic vessel mask derived from the annotated skeleton.

To ensure accurate annotation, the user can toggle between different image layers, such as displaying the skeleton on segmentation or the skeleton on ICA, to verify correct color propagation. By repeating this process for all clinically relevant vessels, any remaining unmarked skeleton segments are semantically labeled using colors selected in ”Select Vessels”, and their corresponding vascular regions are automatically colorized. Finally, the ”Visualize” button is clicked to generate the complete semantic segmentation, integrating labels for all vessels. The resulting semantic mask is stored in both colored format for visualization and MAT format for further processing.

This annotation interface is specifically applied to dataset D_2, as semantic segmentation is inherently more complex than binary segmentation and requires a larger dataset for effective model training. Additionally, semi-supervised semantic segmentation is computationally demanding due to the increased number of classes in the network. Since ICA primarily focuses on three key coronary arteries, the left anterior descending artery (LAD), left circumflex artery (LCX), and right coronary artery (RCA), as discussed in Section 2.3, each view tends to exhibit a consistent overall coronary vascular structure across different patients.

A comprehensive assessment of dataset D_2 revealed that it consists of 154 LCX-focused ICA sequences, 97 LAD-focused ICA sequences, and 72 RCA-focused ICA sequences. Given the manual effort required for semantic annotation (estimated at 15–20 minutes per case), I prioritized the LCX-focused ICA subset, as it provides the largest population for supervised training. For this subset, up to seven clinically significant vessel categories are annotated, depending on the imaging view and the presence of occlusions, as introduced in Section 2.3. These categories are hierarchically organized as follows: LAD, LCX, diagonal branches (Diag.), obtuse marginal arteries (OMs), ramus intermedius (RI), and all other vessels, with corresponding color codes of red, green, yellow, purple, blue, and gray, for visualization respectively, as shown in Fig. 6.2. In cases of vessel overlap, the pixels will be displayed in the color of the higher hierarchical class. This

annotated dataset is designated as D_3, containing $N = 154$ semantically labeled LCX-focused ICA images, with a total of $C = 7$ classes, including the background. This dataset is divided into training, validation, and test sets as follows: $N_{\text{train}} = 103$, $N_{\text{val}} = 26$, and $N_{\text{test}} = 25$.

However, to preliminarily evaluate the contribution of the GNN within the fusion approach for this task, I begin with binary segmentation using the full dataset D_2 but only consider the annotated frame in each sequence ($T = 1$). The dataset splits into $N_{\text{train}} = 173$, $N_{\text{val}} = 82$, and $N_{\text{test}} = 68$, maintaining a total of 323 samples. In this binary segmentation setup, the number of classes is $C = 2$.

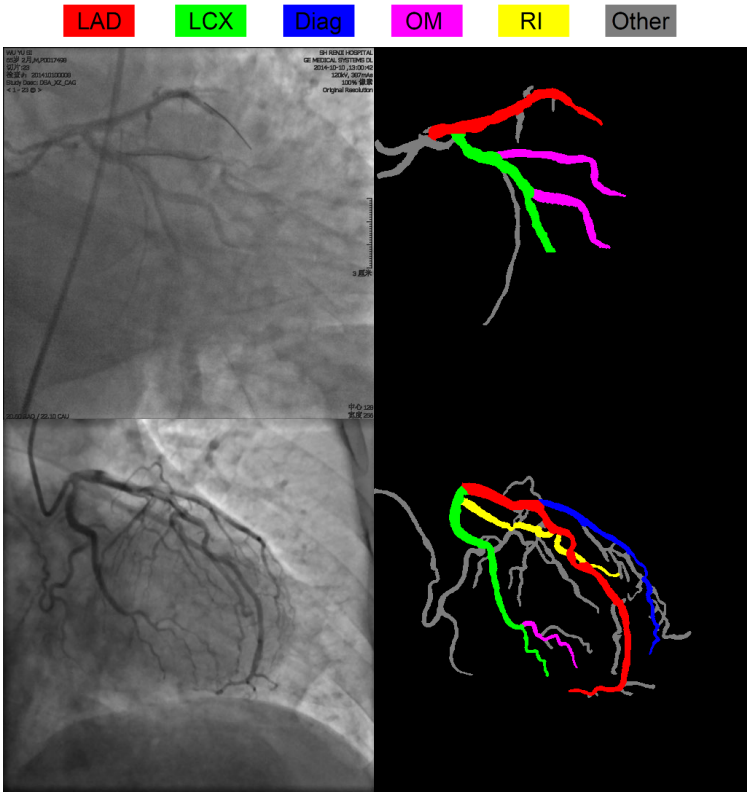


Figure 6.2: Examples of annotated semantic segmentation.

6.2.2 Fusion Approach: CNN + GNN

Integrating vascular structural features derived from a GNN with pixel-wise feature representations learned by a CNN enables a more comprehensive analysis of coronary vessels segmentation. Rather than treating these two models independently, an optimal approach is to fuse the structural insights from the GNN with the spatial and contextual information

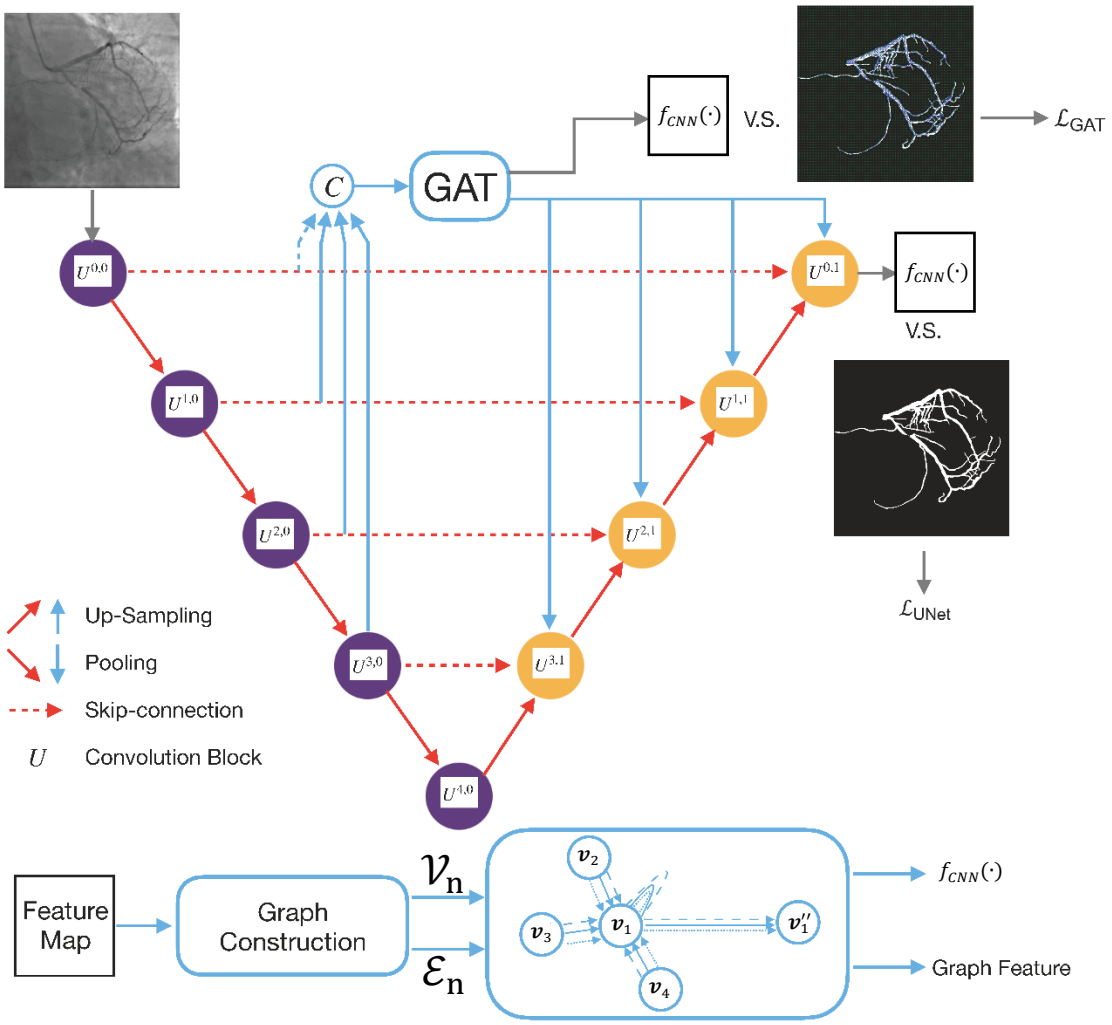


Figure 6.3: The proposed fusion model integrating a GNN and a CNN for vessel delineation from ICA images. The CNN is a 5-layer UNet, with connections between blocks represented by red arrows, while the GNN component is a GAT.

captured by the CNN. This leads to a parallel architecture where both networks operate in tandem, as shown in Fig. 6.3 on binary segmentation as an example, with a graph attention network (GAT) block integrated into four skip connections of the UNet. It takes I_n as input and generates two outputs $f_{CNN}(\cdot)$ and $f_{GNN}(\cdot)$. The entire network is optimized using two loss functions, $\mathcal{L}_{\text{UNet}}$ and $\mathcal{L}_{\text{GAT}}$. The input graph $\mathcal{Q}_n$ is dynamically generated during training using the graph construction method described below.

UNet Setup

The UNet alone is a simple baseline model for image segmentation with VGG convolution blocks fitted. For clarity, in Fig. 6.3, the convolution blocks in the encoder of the UNet are colored purple, and the convolution blocks in the decoder are colored orange. Each block performs sequential operations consisting of convolution, batch normalization, and rectified linear unit (ReLU) activation twice in succession. The loss function $\mathcal{L}_{\text{UNet}}$ is an elastic interaction-based loss function from [109], which has been introduced in detail in Section 4.2.4 and has been proven to preserve vessel connectivity in Chapters 4 and 5. Mathematically, it can be expressed as:

$$\mathcal{L}_{\text{UNet}} = \frac{1}{8\pi} \int_{\mathbb{R}^2} dxdy \int_{\mathbb{R}^2} \frac{\nabla \mathbb{T}(x,y) \cdot \nabla \mathbb{T}(x',y')}{|\vec{r}|} dx'dy' \tag{6.1}$$

where $\mathbb{T}(x,y) = G_s + \alpha\eta_{step}(\phi)$. G_s is the vessel boundary in the gold standard and can be obtained by convolving the ground truth G_n with a 2D Gaussian function. $\eta_{step}(\phi)$ is a regularised Heaviside function that represents the vessel boundary in the generated segmentation. x and y here represent the column and row in the image grid $H \times W$. When $\mathcal{L}_{\text{UNet}}$ is minimized, the boundary in prediction approaches the boundary in the gold standard and closes any disconnected branches by closing the boundary. Given sufficient training resources, UNet can be substituted with alternative skip-connected architectures to enhance performance.

Graph Attention Network

The GAT used in this architecture follows the formulation of Veličković et al. [159], which has been widely applied in spatial graph analysis. GAT introduces a self-attention mechanism to dynamically compute the importance of neighboring nodes, making it more expressive than traditional graph convolutional networks (GCNs). In this model, the node features $\mathcal{V}_n$ and adjacency matrix $\mathcal{E}_n = \mathcal{Z}_n \in \{0,1\}^{|\mathcal{V}_n| \times |\mathcal{V}_n|}$ serve as inputs to GAT, where $\mathcal{Z}_n(r,s) = 1$ if there exists an edge between the r^{th} and s^{th} nodes (i.e., between $\mathbf{v}_r$ and $\mathbf{v}_s$). The network assigns attention scores to neighboring nodes (identified in $\mathcal{Z}_n$) while ensuring that only adjacent nodes contribute to each node's representation. The output of GAT, $f_{GNN}(\cdot)$, represents the probabilities of a node belonging to n classes.

Given a graph $\mathcal{Q}_n = (\mathcal{V}_n, \mathcal{E}_n)$ with $|\mathcal{V}_n|$ nodes, each node is associated with a feature vector $\mathbf{v}_r \in \mathbb{R}^d$. The full node feature matrix is denoted as $\mathcal{V}_n \in \mathbb{R}^{|\mathcal{V}_n| \times d}$. GAT first applies a linear transformation to each node feature as:

$$\mathbf{v}'_r = \theta_w \mathbf{v}_r \tag{6.2}$$

where $\theta_w \in \mathbb{R}^{F' \times F}$ represents a learnable weight matrix. Then, attention scores between connected nodes are computed using an additive attention mechanism:

$$e_{rs} = \text{LeakyReLU}\left(\theta_a^T [\mathbf{v}'_r \parallel \mathbf{v}'_s]\right) \tag{6.3}$$

where $\theta_a \in \mathbb{R}^{2F'}$ is a learnable attention vector and $||$ represents concatenation. LeakyReLU is used here as it provides better performance than ReLU and avoids vanishing gradients. The raw attention scores are then normalized using the softmax function:

$$\text{score}_{rs} = \frac{\exp(e_{rs})}{\sum_{s' \in \mathcal{N}(r)} \exp(e_{rs'})} \tag{6.4}$$

where $\mathcal{N}(r)$ is the set of neighboring nodes. The new feature representation for node $\mathbf{v}_r$ is computed as:

$$\mathbf{v}''_r = \text{activ}\left(\sum_{s \in \mathcal{N}(r)} \text{score}_{rs} \mathbf{v}'_s\right) \tag{6.5}$$

where activ is a nonlinear activation function. Multiple independent attention heads are employed to enhance model stability. Their outputs, denoted as $\mathbf{v}''_r$, are either concatenated at intermediate layers or averaged in the final layer, with the whole process visualized in the blue box of Fig. 6.3.

For GAT to learn from the gold standard node label $\mathcal{A}_n$, a weighted node cross entropy loss is used to take the class imbalance into account:

$$\mathcal{L}_{\text{GAT}} = -\frac{1}{|\mathcal{V}_n|} \sum_{\mathbf{v} \in \mathcal{V}_n} \left[(1-\beta)\, a_{\mathbf{v}} \log \hat{a}_{\mathbf{v}} + \beta(1 - a_{\mathbf{v}}) \log(1 - \hat{a}_{\mathbf{v}})\right] \tag{6.6}$$

where $\beta = \frac{1}{|\mathcal{V}_n|} \sum_{\mathbf{v} \in \mathcal{V}_n} a_{\mathbf{v}}$. Averaging this value calculated for each node will yield the loss for GAT. The graph feature map from the penultimate attention layer of GAT is fused back to the skip connection for the UNet to gain graph information.

Feature Fusion and Graph Construction

With the CNN and GNN architectures introduced separately, the next step is to explore the fusion mechanism and the graph construction process during training to facilitate

effective integration. As illustrated in Fig. 6.3, the feature maps from the four encoder blocks (highlighted in purple) serve a dual purpose: they are passed through the skip connections to the decoder, while also being upsampled to the original input resolution $H \times W$ for graph-based processing. This upsampled feature map is subsequently fed into a fully connected layer, which transforms it into node feature representations, forming the input to the GAT module as $\mathcal{Q}_n = (\mathcal{V}_n, \mathcal{E}_n)$.

On the decoder side, the feature map generated by the GAT module is interpolated back into a spatial representation that matches the resolution of the corresponding CNN feature maps. This interpolated graph-derived feature is then directly concatenated with the skip connection features at multiple scales, enriching the decoder's representation with structural information learned by the GNN. To ensure efficient fusion, the feature alignment between the graph-based and CNN-based representations is crucial. As CNNs inherently operate on structured grids, a grid-like graph structure is adopted during graph construction to maintain spatial consistency with the CNN feature maps, ensuring seamless integration. To achieve this, the graph generation process follows a structured approach inspired by Shin et al. [167], where the feature map derived from the encoder is discretized into a grid-based graph representation.

Node Construction Firstly, the feature map in the original image grid of size $H \times W$ is divided into a smaller sub-grid of size $2^{\chi} \times 2^{\chi}$ as $H = W = 512 = 2^9$ to create $(9/\chi)^2$ number of sub-grids. In every sub-grid, the pixel with the highest value is selected as a node. In all-zero sub-grids, the pixel of the left top corner of the center 2×2 grid is selected as a node. The node's normalized probability is recorded and regarded as the third feature on top of the node location (row and column) of the original image grid. The gold standard of node labels $\mathcal{A}_n$ is determined by extracting the major color from the segmentation gold standard at the grid location.

Edge Construction This edge construction process applies only to nodes that are not derived from an all-zero sub-grid. By selecting a reference node, edges are established between it and all other nodes within a predefined geodesic distance threshold υ. This process is repeated iteratively for all nodes, resulting in a fully constructed graph with

an adjacency matrix generated from the identified edges, with an example of a graph constructed from binary segmentation shown in the top right corner of Fig. 6.3.

6.2.3 Cascaded Approach: CNN + GNN

Unlike the fusion approach, which integrates CNN and GNN in parallel, these models can also be applied in a cascaded manner. As illustrated in Fig. 6.4, the GNN follows the CNN to achieve semantic segmentation of ICA images. This process consists of three sequential stages: ICA binary segmentation using CNN, key node classification using GNN, and node label propagation to the binary segmentation. Each stage operates independently and can be replaced with alternative methods, provided they produce the same format of output.

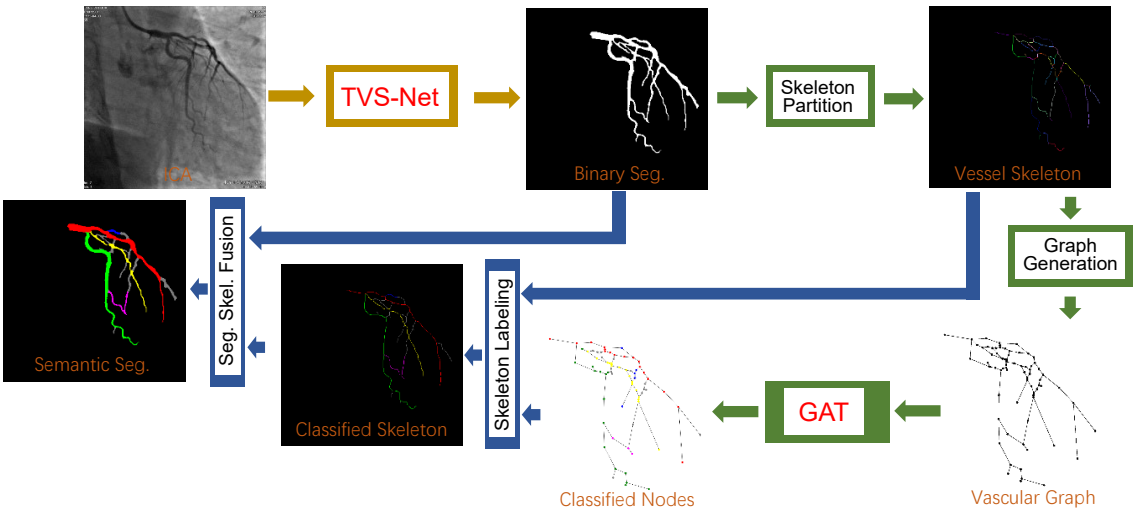


Figure 6.4: The proposed cascaded model combining GNN and CNN for semantic vessel delineation in ICA images. TVS-Net (CNN) is used for binary segmentation (yellow pipeline), GAT (GNN) for node classification (green pipeline), and semantic label propagation is performed in the blue pipeline.

CNN Binary Segmentation

The generation of binary segmentation from ICA using CNN serves as the foundational stage of the entire pipeline. The choice of model is flexible and may include semi-supervised approaches, such as the one introduced in Chapter 4, supervised methods leveraging AutoML frameworks like nnUNet [90], or temporal models incorporating sequential information, such as TVS-Net from Chapter 5. Given its demonstrated superiority over

alternative methods, TVS-Net is selected as the CNN model for this stage. It processes I_n as input and produces $f_{CNN}(\cdot)$ as output, which is subsequently utilized in both the node classification and label propagation stages.

GNN for Node Classification

After achieving accurate binary segmentation of ICA, the segmented binary image undergoes further processing to prepare it for key node classification using GNN. This stage is the most complex and consists of three key steps. First, the binary skeleton is generated from the segmentation output using the skeletonization function $\sigma(\cdot)$, resulting in $\sigma(f_{CNN}(\cdot))$. This skeleton is then partitioned, yielding the partitioned skeleton representation $\sigma_p(f_{CNN}(\cdot))$. Next, a corresponding graph, $\mathcal{Q}_n$, is constructed from the partitioned skeleton, facilitating the representation of vessel structures. However, since grid-structured feature fusion is not required, a more efficient and streamlined graph construction method can be employed. Finally, predicted node classification labels, $f_{GNN}(\cdot)$, are assigned using a specially designed loss function applied to the graph, $\mathcal{Q}_n$, ensuring accurate semantic labeling of key nodes.

Skeleton Partition To begin partitioning the skeleton, I first extract it from the binary segmentation using the method described in Section 5.2.5, producing a one-pixel-wide centerline that preserves essential structural information and enables the identification of bifurcation points. These bifurcation points can be categorized into standard bifurcation points and overlapping points, as illustrated in Fig. 6.5 (a) and (b). In this study, the identification process explicitly includes overlapping points to ensure a more comprehensive skeleton representation.

The skeletonized binary image before partitioning, denoted as $\sigma(f_{CNN}(\cdot))$, consists of pixels with a value of 1 at skeleton locations and 0 elsewhere. To identify bifurcation points, I use a fixed 3×3 connectivity kernel:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

This kernel is convolved with the skeleton image to compute a connectivity map δ, where each value $\delta(x, y)$ indicates the number of active neighboring pixels around location (x, y).

Next, for each pixel where $\delta(x, y) > 2$, I examine the 8-connected neighborhood in a clockwise order to count the number of transitions from 0 to 1. Referring to Fig. 6.5, the binary values of these eight neighbors are denoted by a sequence $[s_1, s_2, ..., s_8]$, and the number of 0-to-1 transitions is calculated as:

$$TsCount(x, y) = \sum_{h=1}^{8} (1 - s_h) \cdot s_{h+1}, \quad \text{with } s_9 = s_1 \tag{6.7}$$

Thus, the bifurcation point mask can be identified as:

$$Mask_B = \{(x, y) \mid TsCount(x, y) \geq 3, \delta(x, y) > 2\} \tag{6.8}$$

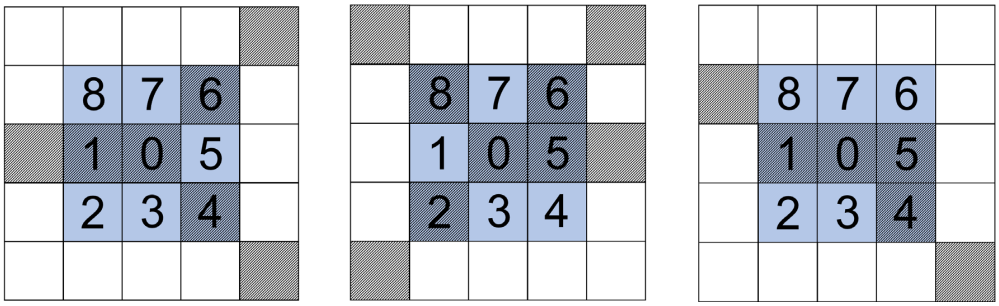


Figure 6.5: (a) Example of a standard bifurcation point. (b) Example of an overlapping point. (c) Example where the degree of connectivity is insufficient for bifurcation point identification. Each case is illustrated using a 5×5 pixel grid, where shaded squares represent skeleton pixels and empty squares represent background pixels. The 3×3 blue square highlights the connectivity kernel and the 0-1 transition region, with numbered positions indicating the sequence of s_h.

The combined criteria of connectivity degree and 0-to-1 transitions help prevent the misidentification of bifurcation points in cases like those shown in Fig. 6.5(c). Such misidentifications often occur when binary-segmented vessels make sharp turns or when skeleton pixels are located near overlapping vessel regions.

However, these criteria may also result in multiple bifurcation points being detected within an overlapping vascular region, as seen in Fig. 6.5(b), where skeleton pixels at positions 0, 5, and 6 are all labeled as bifurcation points. To remove this redundant structural information, adjacent bifurcation points are merged, ensuring that the final bifurcation mask, $Mask_B$, accurately represents the vessel branching structure.

Similarly, by analyzing neighborhood transitions in skeleton pixels, vessel endpoints, where the vessel terminates, can be identified as:

$$Mask_E = \{(x, y) \mid TsCount(x, y) = 1\} \tag{6.9}$$

After detecting bifurcation points and endpoints, the skeleton is partitioned into individual vessel branches $\sigma_p(f_{CNN}(\cdot))$. Each segment begins at an endpoint or bifurcation point and extends until the next bifurcation or termination, with an example of a partitioned skeleton over binary segmentation and corresponding ICA presented in Fig. 6.6. This partitioning process ensures that each vessel segment is uniquely identified, facilitating graph construction for node classification in the GNN pipeline.

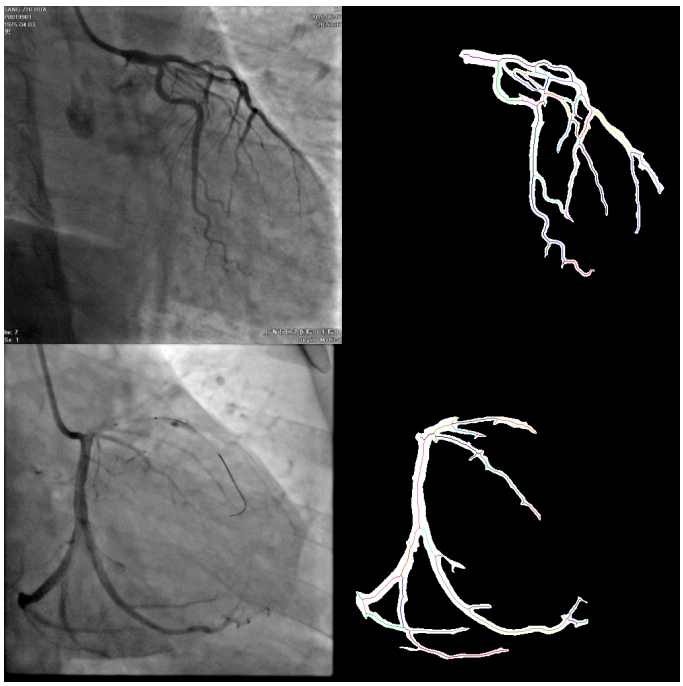


Figure 6.6: Examples of partitioned skeletons on binary segmentation masks from ICA images. The skeleton colors are randomly assigned for visualization purposes only.

For skeleton segments that both begin and end at bifurcation points, an additional midpoint is selected to further refine the partitioning. A midpoint is chosen such that the number of skeleton pixels from the starting point of the segment to this midpoint is equal to the number of skeleton pixels from this midpoint to the ending point of the segment. The collection of all such midpoints forms the mask $Mask_M$.

Graph Generation Unlike the grid-dependent graph generation method used in the CNN-GNN fusion approach, the graph constructed here is optimized to efficiently capture structural information while reducing the complexity of nodes and edges. To achieve this, I directly use the key point masks obtained from skeleton partitioning as node locations, which consist of the union of $Mask_B$, $Mask_E$, and $Mask_M$. While $Mask_B$ (bifurcation points) and $Mask_E$ (endpoints) provide fundamental structural information, incorporating $Mask_M$ (midpoints) further enhances the representation of long, non-terminating branches, particularly those exhibiting curvature, ensuring a more precise structural depiction.

Edges are then established by analyzing the connectivity between key points within the partitioned skeletons. This ensures that each segment between a starting point and an ending point is accurately incorporated into the graph structure as an edge. With the nodes identified and edges established, a preliminary graph is constructed for each binary segmentation. A root node is then selected, corresponding to the bifurcation point where the LCX originates from the LAD, as the LAD is the primary branch on the left side of the heart.

This bifurcation serves as a reference point for constructing an undirected tree graph using the Breadth-First Search (BFS) algorithm, which is subsequently utilized for node feature extraction. Using these two graphs, binary segmentation $f_{CNN}(\cdot)$ and partitioned skeleton $\sigma_p(f_{CNN}(\cdot))$, the final graph $\mathcal{Q}_n$ is constructed, incorporating features indexed from 1 to 37 as detailed in Table 6.1. The 0^{th} feature represents the class of the node $\mathcal{A}_n$, determined by extracting the color from the semantic segmentation gold standard G_n at the node's location (x, y).

The graph structure features extracted from the BFS tree are designed to capture the hierarchical organization of the coronary vessel branching mechanism by embedding the structured relationships, such as LCX originating from LAD, OM branching from LCX, and other vessels following similar patterns. Additionally, the direction and length of skeleton partitions at specific bifurcation points further reflect this structured branching. These features are fundamental in guiding human annotators when assigning semantic labels to vessels, as they closely align with the key characteristics considered during manual annotation.

To ensure a consistent number of features across all nodes, the maximum number

Table 6.1: List of features generated for each node.

Category	No.	Feature
Label	0	Class of the node
Spatial	1-2	Position of the node (x, y)
	3-4	Node relative location to root node
	5-6	Node relative location to image center
	7	Distance to the root node
	8	Distance to graph center
	9	Distance to vascular tree center
	10-13	Distance to each image edge
	14	Number of nodes in 5×5 matrix centered at (x, y)
	15	Number of nodes in 9×9 matrix centered at (x, y)
	16-17	Vessel width and ICA pixel intensity at (x, y)
Graph Structure	18	Node level in the generated BFS tree
	19-20	Max and Min steps to reach one endpoint
	21-22	Number of parent and child nodes
Parent Relations	23-24	Distance to each parent node
	25-26	Direction to each parent node
	27-28	Tortuosity to each parent node
Child Relations	29-31	Distance to each child node
	32-34	Direction to each child node
	35-37	Tortuosity to each child node

of parent and child nodes in the graph is set to 2 and 3, respectively. Each parent node contributes 3 features (distance, direction, and tortuosity), resulting in 6 features for parent relations, while each child node also contributes 3 features, leading to 9 features for child relations. If a node has fewer than 2 parents or 3 children, the missing values are assigned specific placeholders: the distance and tortuosity values are set to -1 since they are inherently non-negative, while the direction is set to 2π to indicate missing values, as valid directions range from $-\pi$ to π. These placeholders ensure the model can distinguish between actual values and missing relationships while maintaining a structured input format. Through this process, a graph representing the skeleton topology is successfully generated for GNN input, with examples visualized in Fig. 6.7.

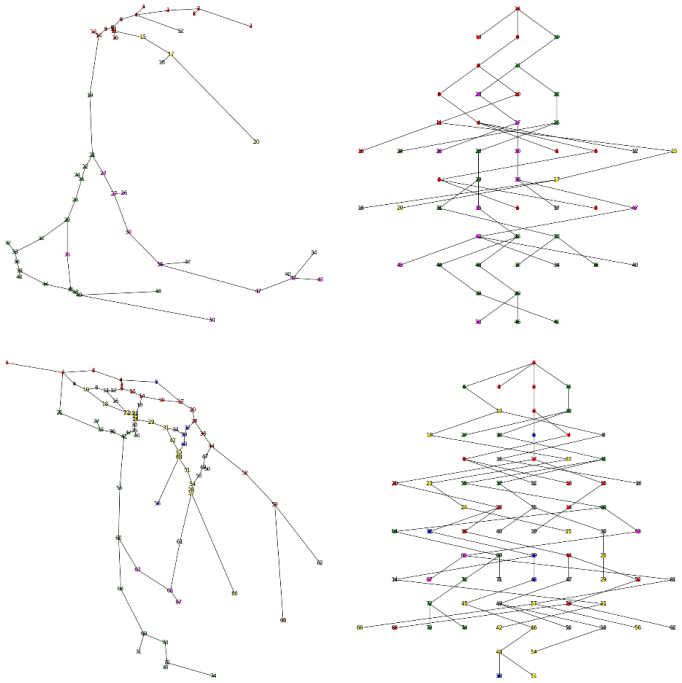


Figure 6.7: Example of a generated graph from a partitioned vessel skeleton and its corresponding tree graph with the colors of the nodes representing their class information.

Graph Attention Network The GNN used in the cascaded approach is a graph attention network (GAT) similar to the one used for the fusion approach in Section 6.2.2. However, as the graph $\mathcal{Q}_n$ generated differs substantially in total node numbers, the input representation is changed from node features and an adjacency matrix to node features and an edge index for each graph. This change enhances memory efficiency, as the edge index format avoids the need for padding sparse and redundant adjacency matrices. By utilizing the edge index, GAT computes distinct attention scores of source and target nodes for addition, rather than concatenating node features as done in Eqs. (6.2) and (6.3). For each node $\mathbf{v}_r$ connected with node $\mathbf{v}_s$, the unnormalized attention score is computed as:

$$e_r = \theta_{a1}^T \mathbf{v}_r \quad \text{and} \quad e_s = \theta_{a2}^T \mathbf{v}_s \tag{6.10}$$

where $\theta_{a1}, \theta_{a2} \in \mathbb{R}^{F'}$ are learnable vectors for the source and target nodes, respectively. The total edge attention score for edge (r, s) is then computed as:

$$e_{rs} = \text{LeakyReLU}(e_r + e_s) \tag{6.11}$$

Then the GAT follows the same process in Eqs. (6.4) and (6.5).

Moreover, as the vascular structure of all the identified vessels differs from that of other vessels, they all exist only in one branch without side branches. Thus, I especially design two penalty loss functions, namely the branch penalty and the propagation penalty, to enforce fewer side branches. The branch penalty penalizes misclassified branching points by considering the impact of incorrect predictions on their connected nodes. Specifically, given the classification probability P_r for node $\mathbf{v}_r$, the misclassification penalty is defined as $1 - P_{r,a_r}$, where a_r denotes the ground truth label of node $\mathbf{v}_r$. To identify branching points, a binary indicator is defined to generate a point set:

$$b_r^{(b)} = \mathbb{1}[\mathsf{deg}(\mathbf{v}_r) > 2] \tag{6.12}$$

where $\mathsf{deg}(\mathbf{v}_r)$ is the node degree, representing the number of neighbors. Finally, the branch misclassification penalty is computed for each edge (r, s), where $\mathbf{v}_s$ is misclassified. The sum of these values yield the branch misclassification loss:

$$\mathcal{L}_{\mathsf{branch}} = \sum_{(r,s)\in\mathcal{E}} b_r^{(b)} \cdot (1 - P_{r,a_r}) \cdot \mathbb{1}[\hat{a}_s \neq a_s] \tag{6.13}$$

where $\mathcal{E}$ is the set of edges in the graph, $\hat{a}_s$ is the predicted class of node $\mathbf{v}_s$, and $\mathbb{1}[\hat{a}_s \neq a_s]$ is an indicator function that equals 1 if node $\mathbf{v}_s$ is misclassified.

The propagation penalty penalizes misclassification errors that propagate downstream from misclassified branching points. Firstly, I identify all misclassified branching points, i.e., nodes satisfying $b_r^{(b)} = 1$ and $\hat{a}_r \neq a_r$. Then, a propagation point set $b_r^{(p)}$ is defined such that $b_r^{(p)} = 1$ for each misclassified node and for all nodes reachable from them via a BFS traversal. The propagation loss is computed as:

$$\mathcal{L}_{\mathsf{prop}} = \sum_{\mathbf{v}_r\in\mathcal{V}} b_r^{(p)} \cdot (1 - P_{r,a_r}) \tag{6.14}$$

To formulate the total loss function, a weight-balanced cross-entropy term is incorporated to enhance classification guidance. Consequently, the final loss function is expressed as:

$$\mathcal{L}_{\mathsf{GAT}} = \mathcal{L}_{\mathsf{CE}} + \lambda_{\mathsf{branch}}\mathcal{L}_{\mathsf{branch}} + \lambda_{\mathsf{prop}}\mathcal{L}_{\mathsf{prop}} \tag{6.15}$$

where $\lambda_{\mathsf{branch}}$ and λ_{prop} are weighting factors that control the relative importance of each penalty term.

Semantic Label Propagation

In the final stage of the semantic vessel segmentation pipeline, a two-step semantic information propagation process is employed. First, the predicted node labels $f_{GNN}(\cdot)$ are propagated to the semantic skeleton. After applying corrections to the semantic skeleton, the semantic information is further propagated to the binary segmentation, producing the final semantic segmentation.

Semantic Skeleton Generation The process of generating a semantic skeleton representation involves recoloring vessel skeleton partitions $\sigma_p(f_{CNN}(\cdot))$ based on the node label prediction of the corresponding graph structure. First, I compute the shortest path distance from the root node to all other nodes. For a partitioned skeleton segment between node $\mathbf{v}_r$ and $\mathbf{v}_s$, the label of the node that is farther from the root node is assigned to the segment. The final semantic skeleton is obtained by recoloring each segment according to its assigned class.

Skeleton Correction The skeleton correction process improves the quality of the semantic skeletons by ensuring they are fully connected and free of artifacts. It begins by grouping skeleton segments that share the same semantic label using connected component labeling. If there are multiple disconnected components within the same class, the method repeatedly identifies the closest pair of components and connects them. For each such pair, it locates the nearest points between the two components and connects them with a new path, which is then assigned the appropriate semantic label. This process continues until all segments within the class are connected into a single structure.

Once connectivity is established, all endpoints in the skeleton with the same semantic label are identified. Among these, only the pair of endpoints connected by the longest path is retained, while all other endpoints and their associated paths are removed. This ensures that only the most structurally relevant and anatomically consistent part of the skeleton is preserved.

Semantic Segmentation Generation After correcting the vessel skeleton, a refined semantic segmentation mask is generated with the semantic label of each skeleton segment mapped to the corresponding vessel segment. The semantic mask is obtained by

expanding each pixel in the semantic skeleton using a disk of increasing radius while maintaining the assigned class label. The expansion stops when the disk touches the vessel boundary for the second time, preventing label leakage outside the vessel region. Only pixels inside the binary segmentation mask are preserved, ensuring alignment with vascular structures. This process is repeated for all skeleton pixels, producing a segmentation mask that integrates the semantic information from the skeleton while maintaining spatial coherence with the vessel boundaries. Similar to the gold standard G_n, the refined semantic segmentation mask also enforces a hierarchical color code, prioritizing the more important vessels:

$$\text{LAD (R)} \succ \text{LCX (G)} \succ \text{Diag. (B)} \succ \text{OM (Magenta)} \succ \text{RI (Yellow)} \succ \text{Other (Gray)}. \quad (6.16)$$

This ordering ensures that primary vessels receive consistent labels before assigning secondary branches.

6.2.4 Training Process

Fusion Approach

The UNet and GAT blocks cannot be trained with a single optimizer, as the loss is unable to back-propagate through the graph construction algorithm. Due to the efficiency of this algorithm, the generation of graph $\mathcal{Q}_n$ is limited to every l iterations. When $\mathcal{Q}_n$ is not updated, GAT is frozen with the old version of $\mathcal{V}_n$ and $\mathcal{E}_n$. After obtaining the semantic segmentation from the network, denoising is applied to remove small pixel clusters. Both semantic segmentation and the gold standard are morphologically skeletonized for evaluation of structural completeness.

Cascaded Approach

The cascaded TVS-Net and GAT are trained separately, with the TVS-Net directly obtained from the model trained in Chapter 5. The training of GAT is more flexible compared to the GAT in the fusion approach due to the use of edge index, resulting in linear complexity rather than quadratic complexity. Similarly, semantic segmentation and the gold standard are both skeletonized for further evaluation.

6.2.5 Evaluation Metrics

To assess the performance of the proposed segmentation method, three conventional macro metrics, namely recall, precision, and Dice coefficient, are used to evaluate both the comparison between the predicted segmentation and its gold standard, as well as between the predicted node semantic labels and their ground truth.

The evaluation of skeletonized binary segmentation results is conducted using two additional macro metrics: completeness C_r and correctness C_p, which represent the buffered versions of recall and precision applied to the skeleton, as detailed in Chapter 5. To extend these metrics from binary to multi-class semantic segmentation, I compute them for each class individually and then take the average across all classes.

6.3 Experimental Results

This section begins by outlining the experimental settings for training both approaches. Since the fusion approach has never been applied to coronary vessels segmentation, I first validate its feasibility on binary segmentation. This initial evaluation also simplifies the process of determining the optimal combination of parameters χ and υ for graph generation. Finally, I provide comprehensive quantitative and qualitative comparisons between the proposed approaches and traditional CNN-based methods.

6.3.1 Experimental Settings

Training of the networks is performed on an NVIDIA V100 GPU. I applied random $90°$, $180°$, or $270°$ rotation with a probability of 0.7 to both the CNN component in two approaches. For binary segmentation of the fusion approach, the Adam optimizer is used for both UNet and GAT, with $\beta_1 = 0.85$ and weight decay $= 10^{-5}$. Learning rate for UNet is set to 10^{-4}, and for GAT is 10^{-5}. The other three key hyperparameters α, χ, and υ are set to 0.35, 4, and 25, respectively. The GAT model consists of a two-layer architecture, with 16 and 8 attention heads in the first and second layers, respectively, and feature dimensions of 3, 16, and 2 across the layers.

For semantic segmentation in the fusion approach, I replace the CNN component from UNet to UNet++ for better spatial feature extraction. The weight decay of the Adam

optimizer is set to $= 10^{-6}$, while the learning rate for the GNN is reduced to $= 5 \times 10^{-6}$, with all other hyperparameters remaining unchanged after a grid search. Since the connectivity-preserving loss function requires applying an argmax operation on the class probability output to determine the predicted boundary for each class, and this operation is not differentiable, the loss function for semantic segmentation in UNet++ is formulated as a weighted cross-entropy loss. In the GAT model, the number of output features in the final layer is increased from 2 to 7, corresponding to six vessel categories and the background.

For semantic segmentation in the cascaded approach, the TVS-Net settings remain unchanged from Chapter 5, as the pre-trained model is directly utilized. Similarly, the GAT model is optimized using the Adam optimizer with $\beta_1 = 0.9$, a weight decay of 10^{-6} and a learning rate of 5×10^{-4} based on extensive grid search. The GAT architecture consists of four layers with 32, 16, 16, and 8 attention heads, respectively, and feature dimensions of 37, 256, 128, 128, and 6 across the layers. The final layer outputs six features, as it excludes the background class. The weighting parameters for the branch and propagation penalties, λ_{branch} and λ_{prop}, are both set to 0.01.

6.3.2 Graph Generation in Fusion Approach

A demonstration of how the values of χ and υ will influence the generated graph is shown in Fig. 6.8. It is clear that no edge is created in the background area, and the number of edges increases with the increase in υ. Additionally, with a higher χ value, the nodes become coarser. Thus, if the connection between nodes clearly depicts the structure of the vascular tree, it is more computationally efficient to use a higher χ with fewer nodes constructed. In Fig. 6.8, graphs created when ($\chi = 3, \upsilon = 15$), ($\chi = 4, \upsilon = 25$), and ($\chi = 4, \upsilon = 35$) are all well presenting the vessel structure, but ($\chi = 4, \upsilon = 25$) yields the least number of nodes and edges, which justifies the selections of these values for the current study.

6.3.3 Binary Segmentation with Fusion Approach

To evaluate the performance of the fusion of GAT into the UNet model for binary segmentation and establish a foundation for semantic segmentation, the proposed framework is trained under two settings: $l = 20$ and $l = 10$. When the value of l is set to a lower

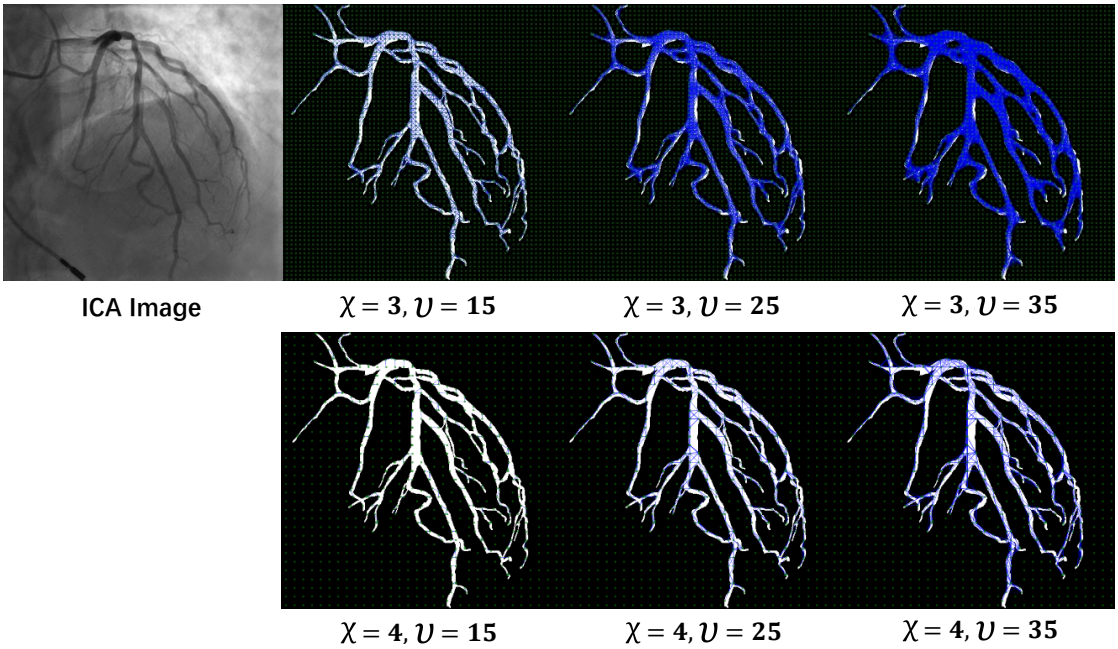


Figure 6.8: Visualization of nodes (green) and edges (blue) constructed using the graph construction method with different parameter values in overlap with G_n. The nodes in the non-vessel area form a regular shape because they are all selected from the same location within their sub-grids. χ is larger than 2, as small numerical values of χ can not extrapolate long-range features for segmentation.

number, it results in more frequent updates to the weights in the GAT model. A baseline model is built using the UNet along with the setup mentioned in Section 6.2.2. The loss function is kept as the connectivity-preserving loss function, the same as the proposed method. The batch size for all models is set to 4 to accommodate the computational constraints of the fusion approach. As reported in Table 6.2, it can be seen that the proposed GAT-UNet fusion model with $l = 10$ achieves the highest performance in terms of the Dice coefficient, recall, and C_r on the test dataset. It surpasses the base UNet by over 5% in terms of recall, implying it misses fewer correct vessels than the UNet and its potential in semantic segmentation over a CNN structure. During training, the GAT-UNet fusion model with $l = 10$ converges its loss before 80 epochs, but UNet takes over 400 epochs for convergence. The precision of GAT-UNet is lower than that of UNet; however, this can be attributed to small or thin vascular structures missing in the gold standard. For example, in the larger white box for the first example in Fig. 6.9, the bottom red arrow points to two distal branches that are not segmented in the gold standard but delineated by

the proposed network. This issue has been elaborated and discussed in Chapter 5. The proposed GAT-UNet fusion model with $l = 10$ captures the most complete vessel structure with more false positives (in blue) that elongate from the end of those true positive vessels (in green).

Table 6.2: Comparison of binary segmentation and skeletonization performance of fusion approach.

Model	Dice (%)	Recall (%)	Precision (%)	$C_r(\%)$	$C_p(\%)$
UNet	79.80±3.46	76.54±6.67	**83.34±4.76**	79.59±7.62	**74.52±6.69**
GAT-UNet (l = 20)	79.93±4.57	79.24±6.88	81.28±5.81	80.55±8.05	73.46±7.68
GAT-UNet (l = 10)	**80.50±4.55**	**81.95±6.94**	79.72±5.99	**81.30±8.10**	72.69±7.56

All values represent mean ± standard deviation (SD).

6.3.4 Semantic Segmentation Results

Building upon the preliminary validation of the fusion approach for binary segmentation in Section 6.3.3, the evaluation is extended to semantic segmentation. The experiments compare three models: a traditional CNN (UNet++), a fusion approach integrating UNet++ and GAT with $l = 10$, and a cascaded approach combining TVS-Net and GAT. As mentioned in Section 6.3.3, the batch size for the fusion approach is restricted to 4 due to GPU memory limitations. To ensure feasibility within the available computational resources, this constraint is applied to all three models in the evaluation. Quantitatively, as shown in Table 6.3, the cascaded approach achieves the highest overall Dice and recall of 63.98% and 66.84%, respectively.

Class-wise, the highest semantic segmentation quality can be found in LCX, with the cascaded model obtaining Dice, recall, and precision of 82.75%, 85.49%, and 82.05%, respectively. The cascaded model also demonstrates a dominating performance in terms of Dice and recall across all classes, apart from the other vessels, with at least 2.27% improvement in Dice and 2.87% improvement in recall. The performance of each model declines as the vessel class moves lower in the hierarchy, with the lowest metrics observed in the RI and other vessel categories. The lower performance metrics for RI can be attributed to its variability as an anatomical feature, unlike the more consistently present LAD and LCX. Since RI may be absent in some patients, its segmentation is inherently more challenging. As shown in Fig. 6.10, where test samples were randomly selected,

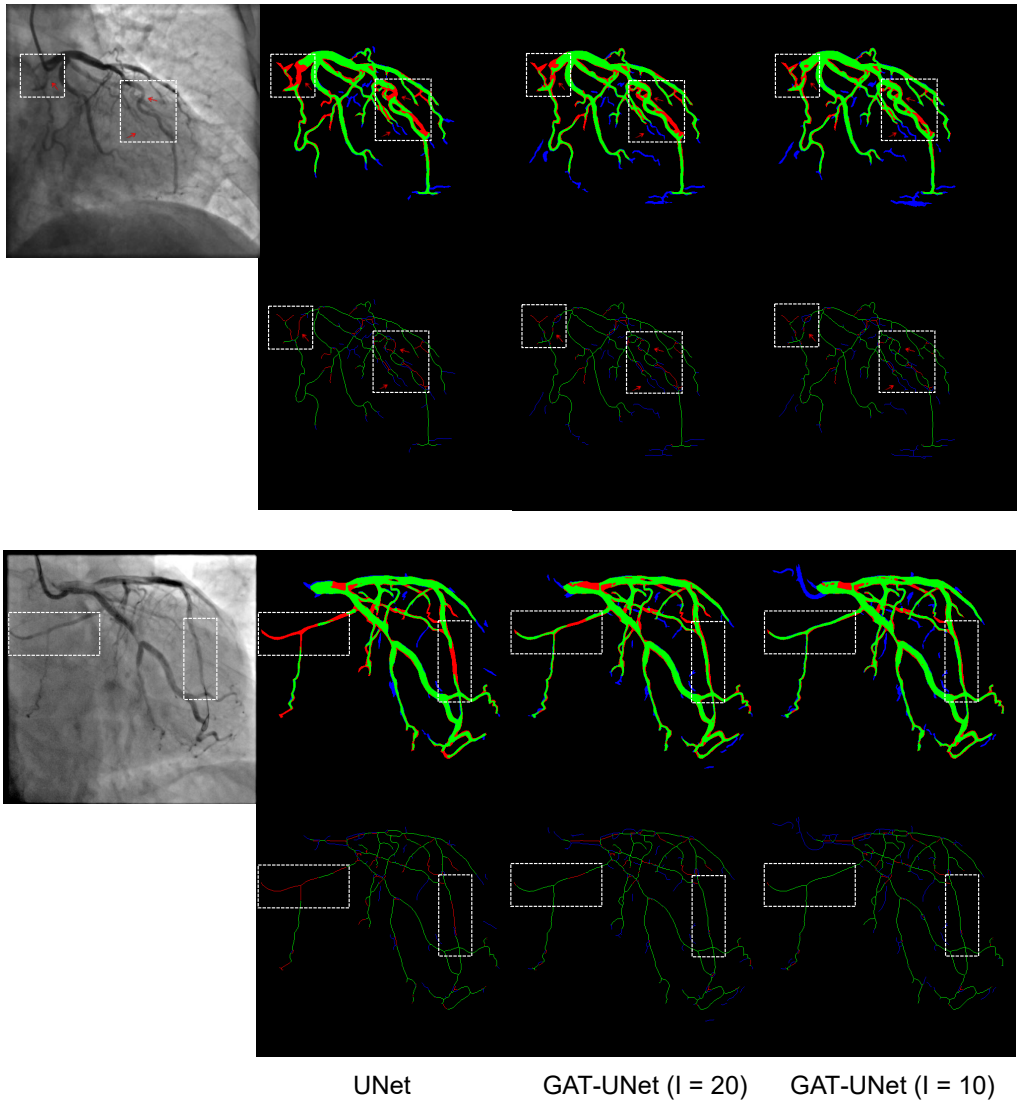


Figure 6.9: Qualitative evaluation of binary segmentation and skeletonization outputs from three different networks. True positive (TP) is shown in green, while false positive (FP) and false negative (FN) are presented in blue and red, respectively. The boxes mark out regions of improved segmetation.

the last sample does not contain a valid RI for segmentation. The reduced performance in segmenting other vessels is primarily due to their low structural consistency within the vascular network. Additionally, these vessels often form thin, overlapping structures in regions enclosed by the LAD and LCX, further complicating their accurate delineation.

However, the cascaded approach demonstrates the lowest performance in segmenting other vessels among the three models. This is primarily due to the largest connected component post-processing step in TVS-Net, which predominantly removes the other vessel class from the binary segmentation output. As a result, these vessels are largely

Table 6.3: Comparison of semantic segmentation performance of CNN, fusion and cascaded approach.

Vessel	Model	Dice (%)	Recall (%)	Precision (%)
LAD	UNet++	80.30±7.81	82.13±13.13	79.83±6.91
	Fusion Model	74.22±10.42	80.70±14.14	70.04±11.07
	Cascaded Model	**82.18±6.12**	**84.49±10.53**	**80.96±6.04**
LCX	UNet++	78.11±11.22	77.57±14.42	80.14±9.23
	Fusion Model	69.01±13.10	74.70±19.89	67.29±12.44
	Cascaded Model	**82.75±11.98**	**85.49±15.72**	**82.05±11.40**
Diag.	UNet++	67.25±19.64	69.60±26.95	**66.48±15.12**
	Fusion Model	60.35±12.56	70.03±20.32	53.21±13.50
	Cascaded Model	**71.87±13.88**	**79.29±23.73**	65.81±16.21
OM	UNet++	60.62±19.19	56.18±25.29	**65.28±16.42**
	Fusion Model	56.35±19.31	58.53±25.11	55.27±20.68
	Cascaded Model	**62.01±20.61**	**68.43±24.99**	58.78±20.43
RI	UNet++	38.54±27.67	34.52±30.33	52.95±26.79
	Fusion Model	36.16±21.82	32.07±22.00	46.22±24.15
	Cascaded Model	**53.45±20.07**	**53.79±23.99**	**53.07±28.61**
Other	UNet++	**52.82±10.63**	56.85±12.79	**51.41±13.13**
	Fusion Model	52.19±8.22	**63.16±8.92**	45.68±9.72
	Cascaded Model	32.79±12.52	29.69±12.62	40.38±15.79
Mean	UNet++	63.08±8.65	64.26±13.65	**67.01±5.70**
	Fusion Model	60.59±6.83	64.75±11.12	57.03±6.42
	Cascaded Model	**63.98±8.11**	**66.84±10.42**	63.51±8.86

All values represent mean ± SD.

absent from the input to the cascaded GAT, making their classification challenging.

Qualitatively in Fig. 6.10, under the same batch size, the cascaded model exhibits the highest structural continuity, with no interruptions along vessels of a specific color, validating the high evaluation metrics values. This improvement is attributed to the skeleton partitioning algorithm and the semantic label propagation method, which effectively preserves vessel connectivity. On the other hand, the fusion model demonstrates the lowest structural continuity, especially on OM in the first sample, LCX in the third sample, and OM in the last sample of Fig. 6.10. This trend is also reflected in the quantitative evaluation of the fusion model. While the fusion model achieves a higher mean recall compared to the conventional UNet++, its significantly lower mean precision leads to an overall lower Dice.

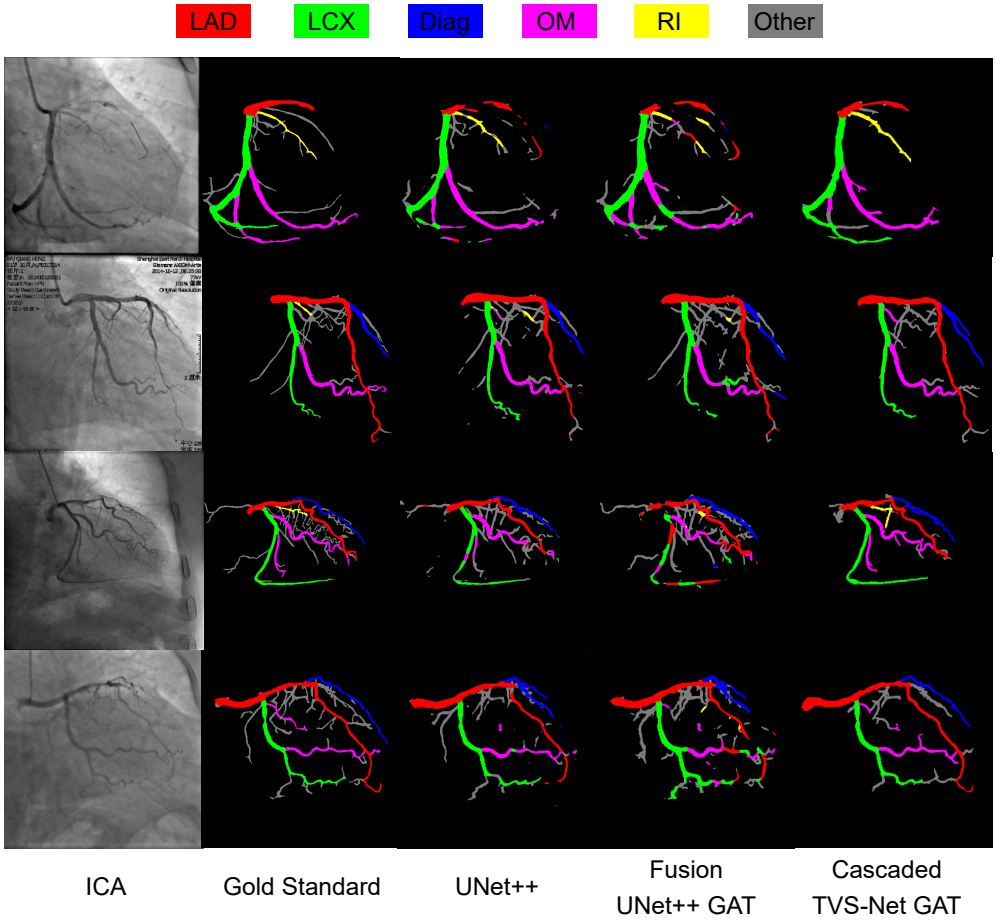


Figure 6.10: Qualitative evaluation of semantic segmentation with the legend indicating the color for different classes.

6.3.5 Efficacy of Penalty Loss

With the cascaded model demonstrating the highest performance in semantic segmentation under a limited batch size, I further investigate its potential by increasing the batch size used in the GAT component. Through grid search, I determine that the optimal performance is achieved with a batch size of 10. Using this batch size, I first evaluate the efficacy of penalty loss functions on node classification results quantitatively in Table 6.4. Next, the quality of semantic segmentation is compared with and without the penalty loss functions against SOTA coronary semantic segmentation method, as presented in Table 6.5.

Table 6.4: Comparison of node classification performance of the cascaded approach.

Model	Dice (%)	Recall (%)	Precision (%)
Cascaded Model w/o $\mathcal{L}_{branch}, \mathcal{L}_{prop}$	81.58±5.44	83.44±7.36	80.67±5.92
Cascaded Model w $\mathcal{L}_{branch}, \mathcal{L}_{prop}$	**84.47±4.28**	**84.70±5.21**	**83.69±4.15**

All values represent mean ± SD.

In Table 6.4, the addition of penalty loss $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ demonstrates the most significant improvements in precision of 3.74%. Qualitatively, the corresponding effects can be found in the 3^{rd} and 5^{th} column of Fig. 6.11. These two contain the corresponding graph prediction from GAT output. The nodes with correct (TP) predictions are illustrated in their corresponding color to the class. The nodes with incorrect (FP or FN) predictions are circled with the color of the correct (TP) prediction. In the 3^{rd} column, where $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ are not used, the circled nodes commonly occur in short side branches. In the 5^{th} column where $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ are in use, the number of circled nodes decreases significantly, especially for the second and third samples. As the number of nodes constructed is between 30 and 60, the correction of several nodes will reflect a significant increase in evaluation metrics.

Table 6.5: Comparison of semantic segmentation between cascaded approaches and SOTA methods.

Vessel	Model	Dice (%)	Recall (%)	Precision (%)
LAD	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**85.21±4.13**	**91.05±6.69**	80.64±6.5
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	84.36±5.71	88.15±9.16	**81.57±6.72**
	Zhao et al. [184]	80.93	N/A	N/A
LCX	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	85.08±11.15	88.37±12.84	80.70±11.14
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**87.05±4.56**	**91.15±8.12**	**83.97±6.20**
	Zhao et al. [184]	75.11	N/A	N/A
Diag.	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	70.89±13.28	79.48±23.25	67.76±18.62
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**71.16±11.07**	**81.62±20.41**	**69.70±15.91**
	Zhao et al. [184]	65.16	N/A	N/A
OM	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	65.18±13.84	74.92±19.32	58.59±17.26
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**72.99±11.00**	**83.25±10.48**	**67.33±14.77**
	Zhao et al. [184]	57.52	N/A	N/A
RI	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	52.42±28.79	50.02±23.97	50.84±22.68
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	57.43±28.02	**57.13±27.20**	**58.07±22.25**
	Zhao et al. [184]	**90.32**	N/A	N/A
Other	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	32.47±11.22	26.31±10.34	46.04±15.94
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**38.69±10.35**	**33.06±13.30**	**51.17±13.85**
Mean	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	65.84±5.29	68.76±7.90	64.32±7.03
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**69.88±4.48**	**71.83±6.74**	**69.40±5.12**
Mean (w/o Other)	Cascaded Model w/o $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	73.04±5.75	78.44±8.79	68.62±8.49
	Cascaded Model w $\mathcal{L}_{\text{branch}}, \mathcal{L}_{\text{prop}}$	**76.08±5.52**	**81.82±7.36**	**73.86±6.80**
	Zhao et al. [184]	73.80	N/A	N/A

All values represent mean or ± SD.

In Table 6.5, penalty loss functions $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ further improve the pixel-wise metrics in all vessel classes apart from LAD. With the accurate delineation of LAD and

LCX in binary segmentation from TVS-Net, the recall for classification of these two vessels reaches around 90%. The most significant improvement can be found in OM vessels, with increases of 11.98% in Dice, 11.12% in recall, and 14.91% in precision. Qualitatively, this can be found in the reduced Magenta side branches in all samples of Fig. 6.11.

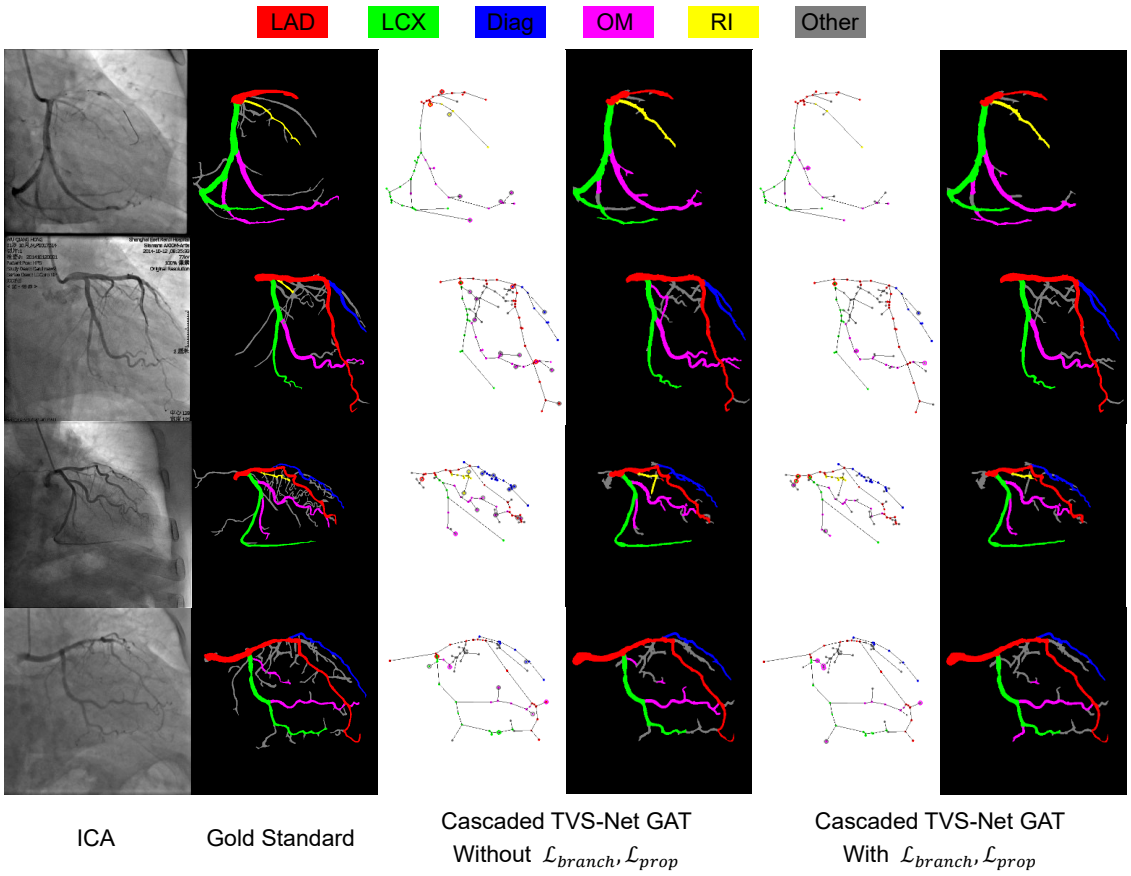


Figure 6.11: Qualitative evaluation of semantic segmentation for cascaded framework alongside the corresponding graph. The node colors in the graph align with the legend, indicating the predicted class. If a node is circled in a different color, it signifies a misclassification.

The state-of-the-art semantic segmentation method introduced by Zhao et al. [184] focuses on 135 LCX-centered images. Its segmentation masks include only the primary branches with no classification for other vessel classes, achieving a Dice score of 89% in the binary segmentation step. Compared to the cascaded model, my approach demonstrates superior identification of LAD, LCX, Diagonal, and OM vessels, with the largest improvement of 26.89% in OM over SOTA method. However, interestingly, SOTA method achieves a Dice score of 90.32% for RI, which remains the most challenging vessel class in my models, as the cascaded model incorporating $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ attains only 57.43%

Dice. With the other vessels excluded, the cascaded model with $\mathcal{L}_{\text{branch}}$ and $\mathcal{L}_{\text{prop}}$ still obtains a higher mean Dice of 76.08% against 73.80% from SOTA method, demonstrating its superiority.

6.3.6 Effectiveness of Skeleton Correction

The skeleton correction method is designed for post-processing where penalty loss functions in the trained GAT model fail to enforce accurate classification. For instance, as illustrated in Fig. 6.12, the RI vessel is incorrectly segmented with multiple side branches. After applying skeleton correction and propagating the refined labels to the segmentation, these side branches are automatically reassigned to the Other class. This refinement enhances the precision of all main vessel classes, leading to an overall increase in precision and Dice, as shown in Table 6.6.

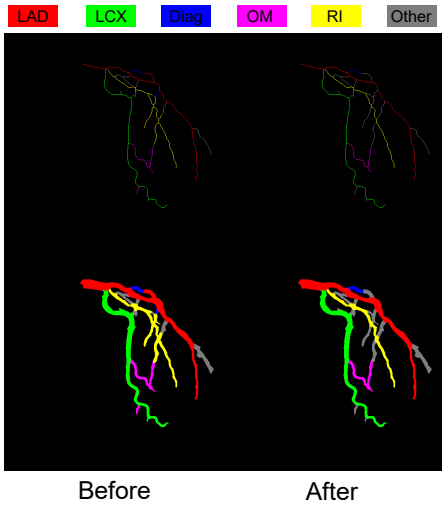


Figure 6.12: Qualitative evaluation of skeleton correction.

Table 6.6: Evaluation of the effectiveness of skeleton correction.

Model	Dice (%)	Recall (%)	Precision (%)
Cascaded Model w/o Skeleton Correction	68.60±4.26	**72.06±6.76**	66.94±5.19
Cascaded Model w Skeleton Correction	**69.88±4.48**	71.83±6.74	**69.40±5.12**

All values represent mean ± SD.

6.4 Discussion and Conclusion

In this chapter, I propose two approaches that integrate CNN and GNN. These approaches complement spatial and structural features in generating semantic coronary vessels segmentation from ICA images. Since no public semantic segmentation dataset is available, a bespoke annotation tool is developed to efficiently add semantic labels to a publicly available binary segmentation dataset D_2. This tool allows users to simply assign semantic labels to binary segmentation skeletons and propagate these labels to the corresponding vessel segments. I select all LCX-focused ICA sequences from D_2 to construct a semantically labeled dataset, D_3, for training and evaluating my proposed approaches, as LCX-focused ICA sequences constitute the largest subset in D_2. The first approach employs a fusion architecture that integrates CNN and GNN in parallel, where structural and spatial features assist each other in generating semantic segmentation. Consequently, graphs representing structural information for GNN analysis are generated on-the-fly from CNN feature maps with a grid-like structure, enabling efficient feature fusion. However, the graph generation process introduces additional computational overhead, leading to increased training time and reduced efficiency during inference. The second approach adopts a cascaded architecture, where CNN and GNN operate in sequence. Initially, spatial features are analyzed by generating binary segmentation through the CNN. Subsequently, structural features are extracted as key points from partitioned binary skeletons, which are then classified by the GNN. The predicted class labels are gradually propagated, first to the skeleton and then to the binary segmentation, resulting in the final semantic segmentation. Hierarchical structural features and misclassification branch penalty loss functions are incorporated to enhance classification. While this approach allows independent tuning of each component, errors in the binary segmentation can carry through the entire pipeline, impacting the accuracy of the final semantic segmentation.

During the evaluation, I first assess the performance of the fusion approach on binary segmentation in D_2. Using UNet as the CNN and GAT as the GNN, strong evidence of improvement is observed, particularly a 7.07% increase in recall, indicating better preservation of vascular structures compared to a conventional UNet with the same connectivity-preserving loss function. This improvement is achieved after selecting appro-

priate values for grid-like graph construction and GAT updates. For semantic segmentation, I replace the baseline model with a conventional UNet++ and also substitute the CNN in the fusion approach with UNet++, while the cascaded approach employs TVS-Net as the CNN and GAT as the GNN. Under a limited batch size, the proposed cascaded model consistently outperforms the conventional UNet++ across all vessel classes except for the "other vessels" category, achieving the highest overall Dice score of 63.98% and recall of 66.84%. This highlights the effectiveness of incorporating structural information to improve classification in vascular structures. However, despite achieving a higher recall compared to conventional UNet++, the fusion model exhibits more fragmentation, with class disconnections leading to a lower Dice score. This issue likely arises because the GAT either inadequately captures structural information due to limited graph input features or because the constrained batch size limits the learning capacity, leading the UNet++ to produce unrealistic false positives. Additionally, since the parallel approach relies on a grid-like graph, it is difficult to impose structural constraints such as maintaining single branches for main vessels and preserving class hierarchy, limiting the potential of the fusion approach.

Without limiting the batch size for GAT in the cascaded model, I fully evaluate its performance against the state-of-the-art method for semantic segmentation. Statistical analyses indicate that my cascaded model significantly improves the segmentation performance of LAD, LCX, Diag., and OM, achieving Dice scores of 84.36%, 87.05%, 71.16%, and 72.99%, respectively. Additionally, the cascaded model attains the highest mean Dice score of 76.08% for the main branches. The incorporation of hierarchical node features from the corresponding tree graph and the application of penalty loss functions with skeleton correction in post-processing enable the proposed cascaded model to analyze structural information and enhance classification in side branches efficiently. When penalty loss functions are not applied, the number of misclassified branches increases, whereas applying penalties to misclassified branches improves the mean Dice score by 3.54% for node classification and by 6.14% for semantic segmentation. The skeleton correction enforces greater connectivity by reducing disconnections and ensuring more consistent branching when propagating node labels to pixels in the binary segmentation. This approach could also be potentially applicable to any model that directly generates semantic segmentation if the segmentation

output is first skeletonized. However, the segmentation of RI presents a challenge, as RI is an anatomical variant that is not present in all patients. Consequently, the cascaded model struggles to achieve accurate RI segmentation. In contrast, the state-of-the-art method attains a high Dice score in RI, which could be attributed to the deliberate inclusion of only patients with an RI in its dataset or differences in annotation protocols. For other vessels, the performance reflects the influence of errors inherited from the binary segmentation. Since some portions of the other vessels are not delineated in the binary segmentation, I can only segment those present in the binary mask. Moreover, classification errors from GAT further lower the evaluation metrics for these vessels. Nevertheless, the cascaded method maintains a high Dice score for all main branches, suggesting that my cascaded approach could perform even better if the binary segmentation mask contained only the main branches, presenting an effective solution for the semantic segmentation of coronary vessels in ICA images.

Chapter 7

Conclusion and Future Works

Chapter contents

7.1 Conclusion

This thesis aims to advance image analysis by introducing various machine learning methods for coronary vessels in invasive coronary angiography (ICA) image data, significantly enhancing automated segmentation and classification capabilities. Given the critical role ICA plays in the diagnosis and management of Coronary Artery Disease (CAD), the advancements presented herein represent meaningful contributions toward improving clinical outcomes through increased accuracy, efficiency, and reliability of image interpretation. The dataset, annotation tools, and methods presented in Chapters 3–6 are clearly defined and practical, enabling downstream tasks to integrate them into existing research workflows seamlessly.

The research began with the introduction of two datasets: the JR dataset D_1 from the

Oxford John Radcliffe (JR) Hospital and the SJTU dataset D_2 from the Renji Hospital of Shanghai Jiao Tong University (SJTU), each designed to fulfill distinct research objectives in this thesis. The JR dataset D_1 includes detailed manual annotations produced using the provided annotation tool. However, it is relatively small in size and contains a large portion of unlabeled data. This setup makes it particularly suited for semi-supervised learning tasks and out-of-distribution inference. In contrast, the SJTU dataset D_2 is substantially larger but has lower segmentation quality, making it an ideal choice for supervised training purposes. Together, these datasets provide the essential foundation for this thesis.

In Chapter 4, I addressed the issue of limited labeled datasets on D_1, a common challenge in medical imaging. Through a semi-supervised learning framework integrating a modified UNet++ architecture and a mean teacher approach, the study demonstrated the ability to leverage both labeled and unlabeled data effectively. UNet++ not only facilitates multi-scale feature extraction but also allows pruning, enabling a more computationally efficient implementation. The mean teacher framework integrates both supervised and consistency losses, encouraging the segmentation predictions to learn from unlabeled images by minimizing discrepancies under perturbations. By introducing an elastic interaction connectivity energy loss function, I significantly improved the structural integrity of binary segmented vessels by forcing the predicted vessel boundaries to approach the true boundaries. I obtained a Dice of 81.7% and a recall of 83.7%, demonstrating its superiority over the state-of-the-art (SOTA) method. Upon further investigation, I observed that reducing the number of labeled samples to 22 led to the most significant performance gains from the semi-supervised method, indicating a potential break-even point for current annotation protocols.

Subsequently, Chapter 5 introduced the Temporal Vessel Segmentation Network (TVS-Net), explicitly designed to extrapolate the temporal information of raw ICA sequences when supervised learning is feasible with dataset D_2. By incorporating temporal context through a combination of 3D and 2D convolutional operations with nested skip-connection, the TVS-Net and its enhanced variant, TVS-Net+, achieved superior segmentation performance. The integration of temporal and spatial information is performed via a fusion block positioned at each skip connection linking the 3D and 2D components of the architecture. With the connectivity energy loss function used for maintaining vascular connectivity and

precision in lesion delineation, TVS-Net achieved superior performance, surpassing six state-of-the-art methods with Dice, recall, and AUPRC scores of 84.31%, 83.41%, and 0.8899, respectively. In subsequent OOD evaluation using dataset D_1, TVS-Net demonstrated superior generalizability, achieving the highest Dice, recall, and AUPRC scores. Additionally, it exhibited improvements of 5.77% and 5.56% in centerline recall and centerline precision over those SOTA methods, respectively. These findings illustrate the critical importance of temporal information in overcoming ambiguities inherent in static-frame analysis. By refining a 10-sample subset of D_2, TVS-Net reveals a 1.2% higher recall compared to the original coarse-grained annotation, demonstrating the feasibility of weak supervision in TVS-Net to even improve on relatively low-quality manual annotation.

Additionally, in Chapter 6, the integration of CNNs and GNNs offered a novel approach to interpreting complex vascular structures for the semantic segmentation of coronary vessels. Supported by semantic annotation D_3 generated using a custom-designed tool from binary segmentation D_2, I proposed two approaches: a fusion model combining CNN and GNN in parallel and a cascaded model integrating CNN and GNN in sequence. Both approaches leverage spatial features extracted by CNN and structural features captured by GNN, with a bespoke vessel graph generation algorithm for feature fusion or label propagation. In the fusion approach, I extract information for grid-like node generation from concatenated CNN feature maps, then integrate the GNN-derived features into the CNN using a grid-like projection. As for the cascaded approach, the CNN first generates a binary segmentation to capture spatial features, after which the GNN classifies key points from the skeleton with a specially designed branching penalty loss and transfers labels to refine the final semantic segmentation. The vascular graph in the cascaded approach is computationally efficient, using the minimum number of nodes to describe the structure while preserving hierarchical information. In the preliminary binary segmentation test with D_2 for the fusion approach before semantic segmentation, the fusion of UNet and GAT achieves faster convergence and better structural preservation, with a Dice score of 80.50% and a recall of 81.95%, compared to a conventional UNet. As for semantic segmentation, statistical analyses show that cascading TVS-Net and GAT significantly improves segmentation performance over conventional UNet++ and the SOTA semantic segmentation method, achieving Dice scores of 84.36%, 87.05%, 71.16%, and 72.99% for

LAD, LCX, Diag., and OM, respectively, while also attaining the highest mean Dice score of 76.08% for the main branches. Thus, the integrated CNN-GNN frameworks provided robust performance, substantially improving conventional CNN segmentation and showcasing GNNs' potential for structural feature analysis in ICA analysis.

In summary, by addressing data scarcity through a semi-supervised framework, refining binary segmentation with temporal information fusion, and enhancing vascular structural analysis through CNN-GNN integration, I conducted a multi-perspective analysis of ICA. These contributions provide potential solutions to key challenges and hold promise for improving the accuracy and reliability of AI-driven cardiovascular imaging technologies in both research and clinical applications for (CAD).

7.2 Future Works

Like many other studies, despite the effectiveness of the proposed approaches, this work has inherent limitations that can guide future research in various directions.

Data-related limitations include the lack of comprehensive patient information in both datasets. Essential variables, such as age, gender, race, smoking, and drinking habits, are missing. This omission introduces the risk of AI bias, particularly sampling bias [196]. For instance, if all the patients involved are over 50 years old, the model may not perform well on younger individuals, resulting in limited generalizability. This bias can be mitigated by curating more representative datasets that include a broader range of demographic and lifestyle factors. In addition, there is potential system and annotation bias, particularly in dataset D_1, which was collected exclusively from a single ICA system. Different manufacturers often employ varying image processing algorithms for visualizing X-ray signals, which means that models trained on a single system may not generalize well to others. Future studies should validate models across multi-institutional and multi-device datasets to ensure broader applicability. Inter-observer variability is another critical limitation [197]. Since neither dataset presents annotations from multiple experts, it is unable to assess the consistency of manual labeling. Although multiple clinicians reviewed annotations in this study, subjective interpretation differences may still impact the training and evaluation of the model. Future work should explore consensus-based or probabilistic

labeling approaches to better capture annotation uncertainty.

To further refine the binary segmentation of ICA images, leveraging the temporal information embedded in raw ICA sequences is crucial. While I previously analyzed short 4-frame ICA sequences, from which dataset D_2 was derived, Hao et al. [34] demonstrated that a 4-frame approach achieves the highest performance. However, I believe that the optimal number of frames should depend on the complexity of the vascular structure and patient-induced motion variations under different conditions. To address this, it would be highly beneficial to create manual annotations based on a single frame from a longer sequence and develop an AutoML-based approach, similar to nnUNet [90], to automatically determine the optimal number of frames for each case. This can be achieved using a multi-path, multi-input segmentation model that dynamically adapts to the specific characteristics of each sequence. Such a strategy may also enhance segmentation performance, as current models occasionally fail unpredictably, limiting their clinical reliability. Ultimately, I aim to achieve accurate semantic segmentation for every frame, enabling precise tracking of diseased vessels for improved visualization and diagnosis.

For semantic segmentation, the results for side branches remain suboptimal. Although the cascaded approach demonstrates better results than the fusion strategy, it still depends on manually selected features for vessel classification. These features may not be sufficiently robust for identifying side branches, even with access to more training data. A promising direction is to incorporate network-learned features by combining CNN and GNN architectures, where features are exchanged effectively to enhance segmentation performance. To ensure generalizability across different annotators, ICA imaging systems, and patient conditions, domain adaptation techniques should also be employed. This is especially important because more complex annotations introduce higher variability. Ensuring consistent and reliable model performance across these variations is essential for successful clinical application. Another promising approach, solely using CNN, is to design a multi-output network that generates multiple binary masks, each isolating a single vessel branch. This would allow the application of specialized binary segmentation loss functions, such as connectivity loss, to improve structural accuracy. However, this method would require an extensive number of annotations per branch, as demonstrated in [198], necessitating a large patient dataset to ensure robustness.

Generally, for any black-boxed algorithm application, explainability is always a problem, especially in the medical field [199]. Potential solutions exist, such as Grad-CAM, which generates visual heatmaps highlighting the regions of the input image that contribute most to the model's decision [200]. This allows clinicians to validate whether the model's focus aligns with known anatomical or pathological features. Incorporating these explainability algorithms into ICA analysis could improve their clinical acceptance by providing insights into the model's reasoning, especially in decision-making scenarios such as coronary stenosis detection and treatment planning.

For downstream tasks in both research and clinical applications, the first consideration is cross-domain adaptation, which enables the segmentation model to be transferred to other vascular imaging modalities, such as CT angiography, or even to structurally similar applications like road segmentation in satellite images. When the target domain is another coronary medical imaging modality, integrating multi-modal imaging after adaptation can enhance vessel assessment by providing complementary information. Once segmentation is completed, stenosis detection and grading can be performed to identify and quantify the severity of arterial narrowing, aiding in CAD diagnosis. For example, by calculating the percentage of luminal narrowing in segmentation, clinicians can grade the severity of stenosis [201]. In more advanced evaluation, algorithms can estimate the fractional flow reserve (FFR) non-invasively from 2D ICA using computational fluid dynamics, assessing the functional significance of stenosis [202, 203]. Furthermore, with multiple views of the same patient's vascular structure, 3D reconstruction for better structural visualization from the angiography, as demonstrated in [5], becomes feasible. This reconstruction enables the extraction of advanced biomarkers, such as through blood flow simulation, allowing for a more detailed functional analysis of coronary circulation. Ultimately, these biomarkers can correlate with patient prognosis and treatment response, such as 3D visualization in real-time during stent placement and 3D operation simulation before bypass surgery, contributing to more personalized and data-driven clinical decision-making.

Bibliography

[1] G. A. Roth, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, *et al.*, “Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017,” *The Lancet*, vol. 392, no. 10159, pp. 1736–1788, 2018.

[2] R. C. Cury, S. Abbara, S. Achenbach, A. Agatston, D. S. Berman, M. J. Budoff, K. E. Dill, J. E. Jacobs, C. D. Maroules, G. D. Rubin, *et al.*, “Coronary artery disease - reporting and data system (CAD-RADS),” *Journal of the American College of Cardiology: Cardiovascular Imaging*, vol. 9, no. 9, pp. 1099–1113, 2016.

[3] V. Kočka, “The coronary angiography–an old-timer in great shape,” *Cor et Vasa*, vol. 57, no. 6, pp. e419–e424, 2015.

[4] P. Garrone, G. Biondi-Zoccai, I. Salvetti, N. Sina, I. Sheiban, P. R. Stella, and P. Agostoni, “Quantitative coronary angiography in the current era: principles and applications,” *Journal of Interventional Cardiology*, vol. 22, no. 6, pp. 527–536, 2009.

[5] A. Banerjee, F. Galassi, E. Zacur, G. L. De Maria, R. P. Choudhury, and V. Grau, “Point-cloud method for automated 3d coronary tree reconstruction from multiple non-simultaneous angiographic projections,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1278–1290, 2019.

[6] S. S. Martin, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, B. B. Gibbs, A. Z. Beaton, A. K. Boehme, *et al.*, “2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association,” *Circulation*, vol. 149, no. 8, pp. e347–e913, 2024.

[7] A. S. Go, D. Mozaffarian, V. L. Roger, E. J. Benjamin, J. D. Berry, M. J. Blaha, S. Dai, E. S. Ford, C. S. Fox, S. Franco, *et al.*, “Executive summary: Heart disease and stroke statistics—2014 update,” *Circulation*, vol. 129, no. 3, pp. 399–410, 2014.

[8] K. Suzuki, “Overview of deep learning in medical imaging,” *Radiological Physics and Technology*, vol. 10, no. 3, pp. 257–273, 2017.

[9] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[10] G. Barbastathis, A. Ozcan, and G. Situ, “On the use of deep learning for computational imaging,” *Optica*, vol. 6, no. 8, pp. 921–943, 2019.

[11] Z. Gao, L. Wang, R. Soroushmehr, A. Wood, J. Gryak, B. Nallamothu, and K. Najarian, “Vessel segmentation for X-ray coronary angiography using ensemble methods with deep learning and filter-based features,” *BMC Medical Imaging*, vol. 22, no. 1, p. 10, 2022.

[12] H. He, A. Banerjee, R. P. Choudhury, and V. Grau, “Deep learning based coronary vessels segmentation in X-ray angiography using temporal information,” *Medical Image Analysis*, vol. 102, p. 103496, 2025.

[13] M. Nobre Menezes, J. L. Silva, B. Silva, T. Rodrigues, C. Guerreiro, J. P. Guedes, M. O. Santos, A. L. Oliveira, and F. J. Pinto, “Coronary X-ray angiography segmentation using artificial intelligence: a multicentric validation study of a deep learning model,” *The International Journal of Cardiovascular Imaging*, vol. 39, no. 7, pp. 1385–1396, 2023.

[14] H. He, A. Banerjee, M. Beetz, R. P. Choudhury, and V. Grau, “Semi-supervised coronary vessels segmentation from invasive coronary angiography with connectivity-preserving loss function,” in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2022.

[15] K. Iyer, B. K. Nallamothu, C. A. Figueroa, and R. R. Nadakuditi, “A multi-stage neural network approach for coronary 3D reconstruction from uncalibrated X-ray angiography images,” *Scientific Reports*, vol. 13, no. 1, p. 17603, 2023.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[17] F.-J. Neumann, M. Sousa-Uva, A. Ahlsson, F. Alfonso, A. P. Banning, U. Benedetto, R. A. Byrne, J.-P. Collet, V. Falk, S. J. Head, *et al.*, “2018 ESC/EACTS guidelines on myocardial revascularization,” *European Heart Journal*, vol. 40, no. 2, pp. 87–165, 2019.

[18] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, *et al.*, “Heart disease and stroke statistics—2022 update: a report from the American Heart Association,” *Circulation*, vol. 145, no. 8, pp. e153–e639, 2022.

[19] D. Sharif, A. Sharif-Rasslan, C. Shahla, A. Khalil, and U. Rosenschein, “Differences in coronary artery blood velocities in the setting of normal coronary angiography and normal stress echocardiography,” *Heart International*, vol. 10, no. 1, pp. heartint–5000221, 2015.

[20] S. D. Fihn, J. M. Gardin, J. Abrams, K. Berra, J. C. Blankenship, A. P. Dallas, P. S. Douglas, J. M. Foody, T. C. Gerber, A. L. Hinderliter, *et al.*, “2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease,” *Circulation*, vol. 126, no. 25, pp. e354–e471, 2012.

[21] P. W. Wilson, R. B. D’Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, “Prediction of coronary heart disease using risk factor categories,” *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[22] A. Hubert, A. Seitz, V. M. Pereyra, R. Bekeredjian, U. Sechtem, and P. Ong, “Coronary artery spasm: the interplay between endothelial dysfunction and vascular

smooth muscle cell hyperreactivity," *European Cardiology Review*, vol. 15, p. e12, 2020.

[23] I. Savulescu-Fiedler, R. O. Baz, R. A. Baz, C. Scheau, and A. Gegiu, "Coronary artery spasm: From physiopathology to diagnosis," *Life*, vol. 15, no. 4, p. 597, 2025.

[24] R. Hajar, "Risk factors for coronary artery disease: historical perspectives," *Heart Views*, vol. 18, no. 3, pp. 109–114, 2017.

[25] J. E. Roeters van Lennep, H. T. Westerveld, D. W. Erkelens, and E. E. van der Wall, "Risk factors for coronary heart disease: implications of gender," *Cardiovascular Research*, vol. 53, no. 3, pp. 538–549, 2002.

[26] M. Zreik, R. W. van Hamersvelt, N. Khalili, J. M. Wolterink, M. Voskuil, M. A. Viergever, T. Leiner, and I. Išgum, "Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1545–1557, 2019.

[27] W. J. Davros, "Fluoroscopy: basic science, optimal use, and patient/operator protection," *Techniques in Regional Anesthesia and Pain Management*, vol. 11, no. 2, pp. 44–54, 2007. Imaging for Interventional Management of Chronic Pain.

[28] T. C. Owens, N. Anton, and M. F. Attia, "CT and X-ray contrast agents: Current clinical challenges and the future of contrast," *Acta Biomaterialia*, vol. 171, pp. 19–36, 2023.

[29] C. Tamburino, *Left Main Coronary Artery Disease: A Practical Guide for the Interventional Cardiologist*. Springer Science & Business Media, 2009.

[30] P. Green, P. Frobisher, and S. Ramcharitar, "Optimal angiographic views for invasive coronary angiography: A guide for trainees," *British Journal of Cardiology*, vol. 23, pp. 110–3, 2016.

[31] M. Dwaik and S. Smerat, "Current standards and potential future advancements in the spatial resolution development of cardiac CT," *Cardiology Vascular Research*, vol. 7, no. 1, pp. 1–5, 2023.

[32] A. Kumar, P. Kumar, and S. Srivastava, “A skewness reformed complex diffusion based unsharp masking for the restoration and enhancement of Poisson noise corrupted mammograms,” *Biomedical Signal Processing and Control*, vol. 73, p. 103421, 2022.

[33] V. Göreke, “A novel method based on Wiener filter for denoising Poisson noise from medical X-Ray images,” *Biomedical Signal Processing and Control*, vol. 79, p. 104031, 2023.

[34] D. Hao, S. Ding, L. Qiu, Y. Lv, B. Fei, Y. Zhu, and B. Qin, “Sequential vessel segmentation via deep channel attention network,” *Neural Networks*, vol. 128, pp. 172–187, 2020.

[35] I. Žuža, T. Nadarević, T. Jakljević, N. Bartolović, and S. Kovačić, “The effect of severe coronary calcification on diagnostic performance of computed tomography-derived fractional flow reserve analyses in people with coronary artery disease,” *Diagnostics*, vol. 14, p. 1738, 08 2024.

[36] A. M. McGuire, C. D. Smith, J. H. Chamberlin, D. Maisuria, A. Tóth, U. J. Schoepf, J. O’Doherty, R. F. Munden, J. Burt, D. Baruah, *et al.*, “Reduction of beam hardening artifact in photon-counting computed tomography: Using low-energy threshold polyenergetic reconstruction,” *Journal of Cardiovascular Computed Tomography*, vol. 17, no. 5, pp. 356–357, 2023.

[37] M. J. Budoff, D. Dowe, J. G. Jollis, M. Gitter, J. Sutherland, E. Halamert, M. Scherer, R. Bellinger, A. Martin, R. Benton, *et al.*, “Diagnostic performance of 64-multidetector row coronary computed tomographic angiography for evaluation of coronary artery stenosis in individuals without known coronary artery disease: Results from the prospective multicenter ACCURACY (assessment by coronary computed tomographic angiography of individuals undergoing invasive coronary angiography) trial,” *Journal of the American College of Cardiology*, vol. 52, no. 21, pp. 1724–1732, 2008.

[38] I. Aly, A. Rizvi, W. Roberts, S. Khalid, M. W. Kassem, S. Salandy, M. du Plessis, R. S. Tubbs, and M. Loukas, “Cardiac ultrasound: An anatomical and clinical review,” *Translational Research in Anatomy*, vol. 22, p. 100083, 2021.

[39] P. S. Rajiah, C. J. François, and T. Leiner, “Cardiac MRI: state of the art,” *Radiology*, vol. 307, no. 3, p. e223008, 2023.

[40] C. R. Sirtori, F. Labombarda, S. Castelnuovo, and R. Perry, “The use of echocardiography for the non-invasive evaluation of coronary artery disease,” *Annals of Medicine*, vol. 49, no. 2, pp. 134–141, 2017.

[41] A. Chiribiri, M. Ishida, E. Nagel, and R. M. Botnar, “Coronary imaging with cardiovascular magnetic resonance: current state of the art,” *Progress in Cardiovascular Diseases*, vol. 54, no. 3, pp. 240–252, 2011.

[42] J. Crowhurst, M. Whitby, D. Thiele, T. Halligan, A. Westerink, S. Crown, and J. Milne, “Radiation dose in coronary angiography and intervention: Initial results from the establishment of a multi-centre diagnostic reference level in Queensland public hospitals,” *Journal of Medical Radiation Sciences*, vol. 61, 09 2014.

[43] E. Topol and R. Califf, *Textbook of Cardiovascular Medicine*. Textbook of Cardiovascular Medicine, Lippincott Williams & Wilkins, 2007.

[44] J. A. Nichols, H. W. Herbert Chan, and M. A. Baker, “Machine learning: applications of artificial intelligence to imaging and diagnosis,” *Biophysical Reviews*, vol. 11, no. 1, pp. 111–118, 2019.

[45] R. K. Mishra, G. S. Reddy, and H. Pathak, “The understanding of deep learning: A comprehensive review,” *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 5548884, 2021.

[46] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, “Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics,” *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 71–91, 2018.

[47] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, “Multiscale vessel enhancement filtering,” in *Medical Image Computing and Computer-Assisted Intervention — MICCAI’98*, pp. 130–137, Springer, 1998.

[48] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, “Detection of blood vessels in retinal images using two-dimensional matched filters,” *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 263–269, 1989.

[49] J. Yang, S. Ma, Q. Sun, W. Tan, M. Xu, N. Chen, and D. Zhao, “Improved Hessian multiscale enhancement filter,” *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 3267–3275, 2014.

[50] C. Kirbas and F. Quek, “A review of vessel extraction techniques and algorithms,” *Association for Computing Machinery Computing Surveys (CSUR)*, vol. 36, no. 2, pp. 81–121, 2004.

[51] M. A. Bhuiyan, B. Nath, and J. Chua, “An adaptive region growing segmentation for blood vessel detection from retinal images,” in *International Conference on Computer Vision Theory and Applications*, vol. 2, pp. 404–409, SCITEPRESS, 2007.

[52] B. Al-Diri, A. Hunter, and D. Steel, “An active contour model for segmenting and measuring retinal vessels,” *IEEE Transactions on Medical Imaging*, vol. 28, no. 9, pp. 1488–1497, 2009.

[53] T. Jerman, F. Pernuš, B. Likar, and Ž. Špiclin, “Enhancement of vascular structures in 3D and 2D angiographic images,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2107–2118, 2016.

[54] C. Wang, R. Moreno, and Ö. Smedby, “Vessel segmentation using implicit model-guided level sets,” in *MICCAI Workshop” 3D Cardiovascular Imaging: a MICCAI Segmentation Challenge”, Nice France, 1st of October 2012.*, 2012.

[55] G. Pizaine, E. D. Angelini, I. Bloch, and S. Makram-Ebeid, “Vessel geometry modeling and segmentation using convolution surfaces and an implicit medial axis,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1421–1424, IEEE, 2011.

[56] S. Cetin, A. Demir, A. Yezzi, M. Degertekin, and G. Unal, “Vessel tractography using an intensity based tensor model with branch detection,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 348–363, 2012.

[57] K. Fang, D.-F. Wang, L. M. Lui, S.-J. Zhou, W. C. Chu, A. T. Ahuja, and P. A. Heng, “3D model-based method for vessel segmentation in TOF-MRA,” in *2011 International Conference on Machine Learning and Cybernetics*, vol. 4, pp. 1607–1611, IEEE, 2011.

[58] R. Socher, A. Barbu, and D. Comaniciu, “A learning based hierarchical model for vessel segmentation,” in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1055–1058, IEEE, 2008.

[59] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[60] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *Journal of the Association for Computing Machinery*, vol. 58, no. 3, pp. 1–37, 2011.

[61] V. Gupta, A. Kale, and H. Sundar, “A robust and accurate approach to automatic blood vessel detection and segmentation from angiography X-ray images using multistage random forests,” in *Medical Imaging 2012: Computer-Aided Diagnosis*, vol. 8315, pp. 704–709, SPIE, 2012.

[62] M. Jin, R. Li, J. Jiang, and B. Qin, “Extracting contrast-filled vessels in X-ray angiography by graduated RPCA with motion coherency constraint,” *Pattern Recognition*, vol. 63, pp. 653–666, 2017.

[63] K. Gurney, *An introduction to neural networks*. Chemical Rubber Company Press, 2018.

[64] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, 1987.

[65] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[68] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd International Conference on Learning Representations (ICLR)*, pp. 1–14, 2015.

[69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[70] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[71] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[72] 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 (CVPR)*, pp. 3431–3440, 2015.

[73] N. Qian, “On the momentum term in gradient descent learning algorithms,” *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.

[74] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[75] X. Tao, H. Dang, X. Zhou, X. Xu, and D. Xiong, “A lightweight network for accurate coronary artery segmentation using X-ray angiograms,” *Frontiers in Public Health*, vol. 10, p. 892418, 2022.

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[77] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[78] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[79] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.

[80] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

[81] T. Dao and A. Gu, “Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality,” in *International Conference on Machine Learning*, pp. 10041–10071, 2024.

[82] H. Zhang, Y. Zhu, D. Wang, L. Zhang, T. Chen, Z. Wang, and Z. Ye, “A survey on visual mamba,” *Applied Sciences*, vol. 14, no. 13, p. 5683, 2024.

[83] W. Yu and X. Wang, “Mambaout: Do we really need mamba for vision?,” *arXiv preprint arXiv:2405.07992*, 2024.

[84] Z. Feng, J. Yang, and L. Yao, “Patch-based fully convolutional neural network with skip connections for retinal blood vessel segmentation,” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1742–1746, 2017.

[85] Z. Feng, J. Yang, L. Yao, Y. Qiao, Q. Yu, and X. Xu, “Deep retinal image segmentation: a FCN-based architecture with short and long skip connections for retinal image segmentation,” in *Neural Information Processing*, pp. 713–722, Springer, 2017.

[86] G. Girish, B. Thakur, S. R. Chowdhury, A. R. Kothari, and J. Rajan, “Segmentation of intra-retinal cysts from optical coherence tomography images using a fully convolutional neural network model,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 296–304, 2018.

[87] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, “Deep retinal image understanding,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016*, pp. 140–148, Springer, 2016.

[88] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer, 2018.

[89] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “UNet 3+: A full-scale connected UNet for medical image segmentation,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, 2020.

[90] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203 – 211, 2020.

[91] Y. Gao, L. Zhang, J. Zhao, and Z. Jiang, “Improved U-Net with channel and spatial attention for coronary angiography segmentation,” in *2022 16th ICME International Conference on Complex Medical Engineering*, pp. 123–126, IEEE, 2022.

[92] J. Zhang, Z. Hua, K. Yan, K. Tian, J. Yao, E. Liu, M. Liu, and X. Han, “Joint fully convolutional and graph convolutional networks for weakly-supervised segmentation of pathology images,” *Medical Image Analysis*, vol. 73, p. 102183, 2021.

[93] Z. Yang, H. Chen, Z. Qian, Y. Zhou, H. Zhang, D. Zhao, B. Wei, and Y. Xu, “ Region Attention Transformer for Medical Image Restoration ,” in *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15007, Springer Nature Switzerland, October 2024.

[94] O. Petit, N. Thome, C. Rambour, and L. Soler, “U-Net Transformer: Self and cross attention for medical image segmentation,” in *Machine Learning in Medical Imaging*, pp. 267–276, 2021.

[95] X. Hu, J. Chen, and Y. Chen, “RegMamba: An improved mamba for medical image registration,” *Electronics*, vol. 13, no. 16, 2024.

[96] J. Ruan and S. Xiang, “VM-UNet: Vision Mamba UNet for medical image segmentation,” *Computer Research Repository*, vol. abs/2402.02491, 2024.

[97] Y. Yue and Z. Li, “MedMamba: Vision mamba for medical image classification,” *arXiv preprint arXiv:2403.03849*, 2024.

[98] M. Simon, E. Rodner, and J. Denzler, “ImageNet pre-trained models with batch normalization,” *arXiv preprint arXiv:1612.01452*, 2016.

[99] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

[100] S. U. Amin and M. S. Hossain, “Edge intelligence and internet of things in healthcare: A survey,” *IEEE Access*, vol. 9, pp. 45–59, 2020.

[101] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[102] B. Dufumier, P. Gori, I. Battaglia, J. Victor, A. Grigis, and E. Duchesnay, “Benchmarking CNN on 3D anatomical brain MRI: architectures, data augmentation and deep ensemble learning,” *arXiv preprint arXiv:2106.01132*, 2021.

[103] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations (ICLR)*, 2014.

[104] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[105] J. Kshirsagar, J. McNulty, B. Taji, D. So, A.-Y. Chong, P. Theriault-Lauzier, A. Wisniewski, and S. Shrimohammadi, “Generative AI-Assisted novel view synthesis of coronary arteries for angiography,” in *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, IEEE, 2024.

[106] S. U. Oh, H. B. Park, H. S. Jeong, J. N. Lee, R. Heo, Y. T. Hong, and H. J. Chang, “Deep learning approaches for image restoration in invasive coronary angiography,” *European Heart Journal*, vol. 45, pp. ehae666–2328, 10 2024.

[107] G. Lin, H. Bai, J. Zhao, Z. Yun, Y. Chen, S. Pang, and Q. Feng, “Improving sensitivity and connectivity of retinal vessel segmentation via error discrimination network,” *Medical Physics*, vol. 49, no. 7, pp. 4494–4507, 2022.

[108] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, “IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3645–3654, 2020.

[109] Y. Lan, Y. Xiang, and L. Zhang, “An elastic interaction-based loss function for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–764, 2020.

[110] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. W. Pluim, U. Bauer, and B. H. Menze, “ClDice - a novel topology-preserving loss function for

tubular structure segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16555–16564, IEEE Computer Society, 2021.

[111] D. Oner, M. Kozinski, L. Citraro, N. C. Dadap, A. G. Konings, and P. Fua, " Promoting Connectivity of Network-Like Structures by Enforcing Region Separation ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 09, pp. 5401–5413, 2022.

[112] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, " A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 12, pp. 8766–8778, 2022.

[113] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[114] Y. Reddy, P. Viswanath, and B. E. Reddy, "Semi-supervised learning: A brief review," *International Journal of Engineering & Technology*, vol. 7, no. 1.8, p. 81, 2018.

[115] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," *Handbook on Neural Information Processing*, pp. 215–239, 2013.

[116] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. *et al.*, eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[117] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[118] S. Agarwal, S. Har-Peled, and D. Roth, "A uniform convergence bound for the area under the ROC curve," in *International Workshop on Artificial Intelligence and Statistics*, pp. 1–8, PMLR, 2005.

[119] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[120] M. Sajjadi, M. Javanmardi, and T. Tasdizen, “Regularization with stochastic transformations and perturbations for deep semi-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[121] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” in *5th International Conference on Learning Representations (ICLR)*, 2017.

[122] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[123] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, “Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, pp. 605–613, Springer, 2019.

[124] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, “Transformation-consistent self-ensembling model for semisupervised medical image segmentation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[125] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin, “Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, pp. 562–571, Springer, 2020.

[126] Q. Chang, Z. Yan, Y. Lou, L. Axel, and D. N. Metaxas, “Soft-label guided semi-supervised learning for bi-ventricle segmentation in cardiac cine MRI,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1752–1755, IEEE, 2020.

[127] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[128] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, “SegAN: Adversarial network with multi-scale L_1 loss for medical image segmentation,” *Neuroinformatics*, vol. 16, pp. 383–392, 2018.

[129] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, “CAT: Constrained adversarial training for anatomically-plausible semi-supervised segmentation,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2146–2161, 2023.

[130] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[131] Y. Skandarani, P.-M. Jodoin, and A. Lalande, “GANs for medical image synthesis: An empirical study,” *Journal of Imaging*, vol. 9, no. 3, p. 69, 2023.

[132] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, “Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11666–11675, 2022.

[133] C. You, R. Zhao, L. H. Staib, and J. S. Duncan, “Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 639–652, Springer, 2022.

[134] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning*, pp. 1597–1607, 2020.

[135] Q. Liu, X. Gu, P. Henderson, and F. Deligianni, “Multi-scale cross contrastive learning for semi-supervised medical image segmentation,” in *British Machine Vision Conference (BMVC)*, 2023.

[136] H. Wang, H. Huang, J. Wu, N. Li, K. Gu, and X. Wu, “Semi-supervised segmentation of cardiac chambers from LGE-CMR using feature consistency awareness,” *BMC Cardiovascular Disorders*, vol. 24, no. 1, p. 571, 2024.

[137] J. Hou, X. Ding, and J. D. Deng, “Semi-supervised semantic segmentation of vessel images using leaking perturbations,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2625–2634, 2022.

[138] A. Tejankar, S. A. Koohpayegani, V. Pillai, P. Favaro, and H. Pirsiavash, “ISD: Self-supervised learning by iterative similarity distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9609–9618, 2021.

[139] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[140] J. A. Fessler, “Tomographic reconstruction using information-weighted spline smoothing,” in *Biennial International Conference on Information Processing in Medical Imaging*, pp. 372–386, Springer, 1993.

[141] D. Atkinson, D. L. Hill, P. N. Stoyle, P. E. Summers, and S. F. Keevil, “An autofocus algorithm for the automatic correction of motion artifacts in MR images,” in *Biennial International Conference on Information Processing in Medical Imaging*, pp. 341–354, Springer, 1997.

[142] Y. Wang, S. M. Resnick, C. Davatzikos, B. L. S. of Aging, and the Alzheimer’s Disease Neuroimaging Initiative, “Analysis of spatio-temporal brain imaging patterns by hidden Markov models and serial MRI images,” *Human Brain Mapping*, vol. 35, no. 9, pp. 4777–4794, 2014.

[143] O. S. Al-Kadi, “Spatio-temporal segmentation in 3-D echocardiographic sequences using fractional brownian motion,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 8, pp. 2286–2296, 2019.

[144] X. Zeng, R. Liao, L. Gu, Y. Xiong, S. Fidler, and R. Urtasun, “DMM-Net: Differentiable mask-matching network for video object segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3929–3938, 2019.

[145] L. Bao, B. Wu, and W. Liu, “CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5977–5986, 2018.

[146] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, “Reinforcement cutting-agent learning for video object segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9080–9089, 2018.

[147] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, “Segment anything,” *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023.

[148] S. Kim, P. Jin, C. Chen, K. Kim, Z. Lyu, H. Ren, S. Kim, Z. Liu, A. Zhong, T. Liu, *et al.*, “MediViSTA: Medical video segmentation via temporal fusion SAM adaptation for echocardiography,” *IEEE Journal of Biomedical and Health Informatics*, 2025.

[149] J. Ollion, M. Maliet, C. Giuglaris, E. Vacher, and M. Deforet, “DiSTNet2D: Leveraging long-range temporal information for efficient segmentation and tracking,” *PRX Life*, vol. 2, p. 023004, 2024.

[150] R. Su, P. M. van der Sluijs, Y. Chen, S. Cornelissen, R. van den Broek, W. H. van Zwam, A. van der Lugt, W. J. Niessen, D. Ruijters, and T. van Walsum, “CAVE: Cerebral artery–vein segmentation in digital subtraction angiography,” *Computerized Medical Imaging and Graphics*, vol. 115, p. 102392, 2024.

[151] Q. Xie, M. Guo, L. Mou, D. Zhang, D. Chen, C. Shan, Y. Zhao, R. Su, and J. Zhang, “DSCA: A Digital subtraction angiography sequence dataset and spatio-temporal model for cerebral artery segmentation,” *IEEE Transactions on Medical Imaging*, pp. 1–1, 2025.

[152] A. Vlontzos and K. Mikolajczyk, “Deep segmentation and registration in X-ray angiography video,” in *British Machine Vision Conference (BMVC)*, 2018.

[153] D. Liang, L. Wang, D. Han, J. Qiu, X. Yin, Z. Yang, J. Xing, J. Dong, and Z. Ma, “Semi 3D-TENet: Semi 3D network based on temporal information extraction for coronary

artery segmentation from angiography video," *Biomedical Signal Processing and Control*, vol. 69, p. 102894, 2021.

[154] T. Wan, J. Chen, Z. Zhang, D. Li, and Z. Qin, "Automatic vessel segmentation in X-ray angiogram using spatio-temporal fully-convolutional neural network," *Biomedical Signal Processing and Control*, vol. 68, p. 102646, 2021.

[155] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[156] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[157] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[158] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[159] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[160] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, "Distance metric learning using graph convolutional networks: Application to functional brain networks," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017*, pp. 469–477, Springer, 2017.

[161] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease," *Medical Image Analysis*, vol. 48, pp. 117–130, 2018.

[162] M. Ghorbani, A. Kazi, M. S. Baghshah, H. R. Rabiee, and N. Navab, "RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data," *Medical Image Analysis*, vol. 75, p. 102272, 2022.

[163] J. Jiang, X. Chen, G. Tian, and Y. Liu, "ViG-UNet: vision graph neural networks for medical image segmentation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2023.

[164] P. He, A. Qu, S. Xiao, and M. Ding, "A GNN-based network for tissue semantic segmentation in histopathology image," in *Journal of Physics: Conference Series*, vol. 2504, p. 012047, 2023.

[165] R. Li, Y.-J. Huang, H. Chen, X. Liu, Y. Yu, D. Qian, and L. Wang, "3D graph-connectivity constrained network for hepatic vessel segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1251–1262, 2021.

[166] H. Xu and Y. Wu, "G2vit: Graph neural network-guided vision transformer enhanced network for retinal vessel and coronary angiograph segmentation," *Neural Networks*, vol. 176, p. 106356, 2024.

[167] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, "Deep vessel segmentation by learning graphical connectivity," *Medical Image Analysis*, vol. 58, p. 101556, 2019.

[168] H. Yu, J. Zhao, and L. Zhang, "Vessel segmentation via link prediction of graph neural networks," in *International Workshop on Multiscale Multimodal Medical Imaging*, pp. 34–43, 2022.

[169] D. Zhang, S. Liu, S. Chaganti, E. Gibson, Z. Xu, S. Grbic, W. Cai, and D. Comaniciu, "Graph attention network based pruning for reconstructing 3D liver vessel morphology from contrasted CT images," *arXiv preprint arXiv:2003.07999*, 2020.

[170] M. Pham, S. Alse, C. A. Knoblock, and P. Szekely, "Semantic labeling: a domain-independent approach," in *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference*, pp. 446–462, Springer, 2016.

[171] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[172] G. Holste, R. Sullivan, M. Bindschadler, N. Nagy, and A. Alessio, "Multi-class semantic segmentation of pediatric chest radiographs," in *Medical Imaging 2020: Image Processing*, vol. 11313, pp. 323–330, 2020.

[173] P. Ahmad, H. Jin, R. Alroobaea, S. Qamar, R. Zheng, F. Alnajjar, and F. Aboudi, "MH UNet: A multi-scale hierarchical based architecture for medical image segmentation," *IEEE Access*, vol. 9, pp. 148384–148408, 2021.

[174] Y. Qin, K. Kamnitsas, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi, and A. Nori, "Autofocus layer for semantic segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*, pp. 603–611, Springer, 2018.

[175] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[176] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "SCAN: Structure correcting adversarial network for organ segmentation in chest X-rays," in *International Workshop on Deep Learning in Medical Image Analysis*, pp. 263–273, Springer, 2018.

[177] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel, "A conditional adversarial network for semantic segmentation of brain tumor," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017*, pp. 241–252, Springer, 2018.

[178] Q. Chen, J. Peng, S. Zhao, and W. Liu, "Automatic artery/vein classification methods for retinal blood vessel: A review," *Computerized Medical Imaging and Graphics*, p. 102355, 2024.

[179] A. E. Chowdhury, G. Mann, W. H. Morgan, A. Vukmirovic, A. Mehnert, and F. Sohel, "MSGANet-RAV: A multiscale guided attention network for artery-vein segmentation and classification from optic disc and retinal images," *Journal of Optometry*, vol. 15, pp. S58–S69, 2022.

[180] W. Chen, S. Yu, K. Ma, W. Ji, C. Bian, C. Chu, L. Shen, and Y. Zheng, "TW-GAN: Topology and width aware GAN for retinal artery/vein classification," *Medical Image Analysis*, vol. 77, p. 102340, 2022.

[181] K. J. Noh, S. J. Park, and S. Lee, "Combining fundus images and fluorescein angiography for artery/vein classification using the hierarchical vessel graph network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 595–605, Springer, 2020.

[182] S. Mishra, Y. X. Wang, C. C. Wei, D. Z. Chen, and X. S. Hu, "VTG-net: a CNN based vessel topology graph network for retinal artery/vein classification," *Frontiers in Medicine*, vol. 8, p. 750396, 2021.

[183] X. Xu, P. Yang, H. Wang, Z. Xiao, G. Xing, X. Zhang, W. Wang, F. Xu, J. Zhang, and J. Lei, "AV-casNet: fully automatic arteriole-venule segmentation and differentiation in OCT angiography," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 481–492, 2022.

[184] C. Zhao, R. Bober, H. Tang, J. Tang, M. Dong, C. Zhang, Z. He, Y.-P. Wang, H.-W. Deng, M. L. Esposito, *et al.*, "Semantic segmentation to extract coronary arteries in invasive coronary angiograms," *medRxiv*, pp. 2020–05, 2020.

[185] C. Zhao, Z. Xu, G.-U. Hung, and W. Zhou, "EAGMN: Coronary artery semantic labeling using edge attention graph matching network," *Computers in Biology and Medicine*, vol. 166, p. 107469, 2023.

[186] R. O. Mada, P. Lysyansky, A. M. Daraban, J. Duchenne, and J.-U. Voigt, "How to define end-diastole and end-systole?: Impact of timing on strain measurements," *The Journal of the American College of Cardiology: Cardiovascular Imaging*, vol. 8, no. 2, pp. 148–157, 2015.

[187] T. Cerciello, P. Bifulco, M. Cesarelli, and A. Fratini, “A comparison of denoising methods for X-ray fluoroscopic images,” *Biomedical Signal Processing and Control*, vol. 7, no. 6, pp. 550–559, 2012.

[188] Y. Xiang, A. C. Chung, and J. Ye, “An active contour model for image segmentation based on elastic interaction,” *Journal of Computational Physics*, vol. 219, no. 1, pp. 455–476, 2006.

[189] J. Zhang, R. Gu, G. Wang, H. Xie, and L. Gu, “SS-CADA: A semi-supervised cross-anatomy domain adaptation for coronary artery segmentation,” *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1227–1231, 2021.

[190] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 562–570, PMLR, 2015.

[191] T. Y. Zhang and C. Y. Suen, “A fast parallel algorithm for thinning digital patterns,” *Communications of the Association for Computing Machinery*, vol. 27, no. 3, pp. 236–239, 1984.

[192] R. Youssef, A. Ricordeau, S. Sevestre-Ghalila, and A. Benazza-Benyahya, “Evaluation protocol of skeletonization applied to grayscale curvilinear structures,” in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6, 2015.

[193] X. Chen, L. Yao, and Y. Zhang, “Residual attention U-Net for automated multi-class segmentation of covid-19 chest CT images,” *arXiv preprint arXiv:2004.05645*, 2020.

[194] Y. Qiu, Z. Li, Y. Wang, P. Dong, D. Wu, X. Yang, Q. Hong, and D. Shen, “CorSegRec: A topology-preserving scheme for extracting fully-connected coronary arteries from CT angiography,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 670–680, 2023.

[195] P. E. Hart, N. J. Nilsson, and B. Raphael, “A formal basis for the heuristic determination of minimum cost paths,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[196] J. W. Gichoya, K. Thomas, L. A. Celi, N. Safdar, I. Banerjee, J. D. Banja, L. Seyyed-Kalantari, H. Trivedi, and S. Purkayastha, “AI pitfalls and what not to do: mitigating bias in AI,” *The British Journal of Radiology*, vol. 96, no. 1150, p. 20230023, 2023.

[197] B. Winkfield, C. Aubé, P. Burtin, and P. Calès, “Inter-observer and intra-observer variability in hepatology,” *European Journal of Gastroenterology & Hepatology*, vol. 15, no. 9, pp. 959–966, 2003.

[198] Z. Xian, X. Wang, S. Yan, D. Yang, J. Chen, and C. Peng, “Main coronary vessel segmentation using deep learning in smart medical,” *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 8858344, 2020.

[199] J. Duell, X. Fan, B. Burnett, G. Aarts, and S.-M. Zhou, “A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, IEEE, 2021.

[200] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[201] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M. de Graaf, M. Meijs, *et al.*, “Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography,” *Medical Image Analysis*, vol. 17, no. 8, pp. 859–876, 2013.

[202] G. Yang, L. Li, X. Peng, G. Tang, N. Zheng, Y. Zhao, H. Li, H. Zhang, F. Sun, and H. Ai, “Accuracy and reproducibility of coronary angiography-derived fractional flow reserve in the assessment of coronary lesion severity,” *International Journal of General Medicine*, pp. 3805–3814, 2023.

[203] Z. Zhang, M. Xie, X. Dai, Z. Duan, Z. Lu, L. Cai, R. Gu, L. Shen, Z. Xu, W. Yao, *et al.*, “The prognostic value and economic benefits of coronary angiography-derived

fractional flow reserve-guided strategy in patients with coronary artery disease," *Heliyon*, vol. 9, no. 6, p. e17464, 2023.