People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Ghardaia
Faculty of Science and Technology
Department of Mathematics and Computer Science
Mathematics and Applied Sciences Laboratory

Master Thesis

Presented to obtain the academic Master diploma in Computer Science

Specialty: Intelligent Systems for Knowledge Extraction

A multi source/task deep learning system for non-invasive detection and estimation of coronary artery stenosis

---

Presented by

Hocine Harrouzi & Houdaifa Yahia-cherif

Defended before the jury Members:

Slimane Bellaouar	MCA	Univ. Ghardaia	President
Youcef Mahdjoub	MAA	Univ. Ghardaia	Examiner
Slimane Oulad-Naoui	MCA	Univ. Ghardaia	Supervisor
Abderrahmane Cherif	PhD Student	Univ. Ghardaia	Co-Supervisor

University Year: 2025/2026

Acknowledgement

First and foremost, we express our deepest gratitude to the Almighty God, whose guidance and strength enabled us to complete this work.

We extend heartfelt thanks to our supervisor, Mr. Slimane Oulad-Naoui, and our co-supervisor, Mr. Abderrahmane Cherif, for their invaluable guidance, unwavering support, and expertise that significantly enriched our research. Their insightful direction and contributions to our project were instrumental in its completion.

We are particularly grateful to Oasis Clinic. We would like to thank the administration and staff for their collaboration and for providing the professional environment that allowed us to bridge the gap between academic theory and practical implementation.

All our gratitude goes to the Mathematics and Applied Sciences Laboratory at the University of Ghardaia for their kind assistance and for providing access to essential laboratory resources.

We are also grateful to our families for their unwavering support and to the honourable members of the jury for their time and interest in evaluating this work.

Lastly, we extend our thanks to all who have assisted us directly or indirectly. Their contributions have been invaluable.

Dedication

I dedicate this modest work to

My Beloved Parents, for their endless sacrifices, prayers, and for being the foundation of every success I achieve. May God grant you health and long life.

My Wife, for her patience, understanding, and for being my source of strength and tranquility during the most demanding times of this journey.

My Dear Brothers and Sister, who have always stood by me with encouragement and pride.

My Colleagues at Oasis Clinic, for their support and for fostering a professional atmosphere that inspires continuous learning and growth.

My Partner, Houdaifa Yahia-cherif, for the tireless effort, the shared challenges, and the successful collaboration that brought this project to life.

Hocine Harrouzi

I dedicate this modest work to

My Dear Parents, for their unwavering love, support, and for believing in me when I needed it most. Your guidance is my light.

My Brothers and Sisters, for their constant encouragement and for always being there to share the joy of my progress.

My Friends and Peers, for the unforgettable memories and the support that made this academic path more meaningful.

The Team at Oasis Clinic, for their hospitality and cooperation during the realisation of this project.

My Partner, Hocine Harrouzi, for his professionalism, hard work, and the spirit of brotherhood we shared throughout this project.

Houdaifa Yahia-cherif

الملخص

تعد أمراض القلب والأوعية الدموية السبب الأول للوفيات على مستوى العالم. ويعتبر تصوير الأوعية التاجية الغازي المعيار الذهبي لتشخيص أمراض الشرايين التاجية، إلا أن التقييم اليدوي لهذه الصور لا يزال يتطلب جهداً مكثفاً ويخضع لتباين بين الأطباء في التقييم. تعالج معظم حلول التعلم العميق الحالية مهام تجزئة الأوعية الدموية، واكتشاف التضيق، والقياس الكمي، كنماذج منفصلة يتم تدريبها بشكل تسلسلي. ورغم ظهور مقاربات التعلم متعدد المهام في التصوير المقطعي المحوسب للأوعية التاجية، إلا أن النماذج الشاملة التي تستخرج بشكل مشترك خرائط تجزئة الأوعية الدموية، والتحديد الدقيق لأماكن التضيق، والقياسات الكمية المستمرة للشرايين التاجية من خلال التصوير التداخلي بالأشعة السينية مع معايرة عدم اليقين، ودمج بيانات من وسائط غير تصويرية، وتوفير نطاقات ثقة لا تزال غير مدروسة بشكل كافٍ.

تستقصي هذه الأطروحة ثلاث معماريات متدرجة التعقيد، يُطلق عليها اسم «المسارات»، تم تطويرها وتدريبها على مجموعة بيانات تضم 200 مريضاً. يربط المسار الأول ثلاثة نماذج مدرَّبة بشكل مستقل — نموذج U-Net لتجزئة الأوعية الدموية، ونموذج اكتشاف التضيق المعتمد على معمارية EfficientNet-B4+MAnet ومرحلة انحدار لاستخراج القياسات الكمية للشرايين التاجية — لتعمل بشكل تسلسلي كنموذج مرجعي يبرز القصور المتمثل في تراكم الأخطاء في النهج السائد لتحليل صور التصوير التداخلي.

يقترح المسار الثاني معمارية GCT-Net وهو نموذج متعدد المهام أحادي المصدر يعتمد على مشفر هجين يجمع بين الشبكات العصبية الالتفافية والمحولات. يغذي هذا المشفر ثلاثة رؤوس مخصصة للمهام — جهازي فك تشفير متوازيين لتجزئة الأوعية الدموية، وآلية تجميع مُدركة لمناطق الاهتمام لمواضع التضيق تُستخدم لربط عملية الانحدار الكمي بخريطة التضيق المتوقَّعة — مع استخدام آلية إيقاف انتشار التدرج لحماية خريطة تحديد المواقع، وتطبيق أوزان مبنية على عدم اليقين المتجانس لموازنة دوال الخسارة للمهام المختلفة.

أما المسار الثالث، فيوسّع معمارية GCT-Net لتصبح متعددة المصادر من خلال دمج فرع لبيانات تخطيط كهربية القلب المجدولة، وحقول مستخرجة من تقارير الإجراءات الطبية عبر طبقة عنق زجاجة للتكييف من نوع FiLM تحت إشراف جزئي على البيانات. ويضيف هذا المسار أربعة رؤوس مشتقة من التقارير، ومتنبئات امتثالية مجزَّأة لإنتاج تنبؤات معايَرة للتحليل الكمي الشرياني التاجي.

تقيس المقارنة بين هذه المسارات مدى الفائدة المرجوة من الاستمثال المشترك، حيث رفع المسار الثاني معامل دايس للتشابه لاكتشاف التضيق من 0.465 إلى 0.610، وخفّض نسبة الخطأ في تقدير نسبة تضيق القطر وطول الجزء المتضيق من 18.9 نقطة مئوية و5.7 ملم إلى 11.73 نقطة مئوية و4.51 ملم في تمريرة أمامية واحدة. في المقابل، قايض المسار الثالث دقة التصوير بتقديم مخرجات تفسيرية أشمل. أُجريت المصادقة الخارجية لفرع اكتشاف التضيق باستخدام مجموعة البيانات المرجعية ARCADE حيث حقق الضبط الدقيق للنموذج معامل دايس بلغ 0.523 على 300 صورة اختبارية جديدة.

وأخيراً، تم توظيف النظام ضمن تطبيق ويب يقوم بمعالجة الأرشيفات الطبية الخام بصيغة DICOM وتحويلها إلى تقرير مهيكل للتحليل الكمي الشرياني التاجي.

الكلمات المفتاحية: أمراض الشرايين التاجية، التعلم العميق، التعلم متعدد المهام، دمج المصادر المتعددة، الشبكات الهجينة، التحليل الكمي الشرياني التاجي، التنبؤ الامتثالي، المصادقة الخارجية.

Abstract

Cardiovascular disease is the leading cause of death worldwide, and coronary artery disease is its most common form. To diagnose it, clinicians rely on X-ray coronary angiography, but interpreting these images manually is slow and prone to disagreement between experts. Most deep-learning tools address the three sub-tasks (outlining the vessels, finding the narrowing or stenosis, and measuring its severity through quantitative coronary analysis, QCA) with separate models trained one after another. A few combined models exist for CT scans, but for X-ray angiography no single network yet produces all three outputs at once, fuses additional non-image data, and reports its own confidence. This thesis builds and compares three designs of increasing ambition, termed *flows*, on a 200-patient dataset. Flow 1 is a baseline in which three separate models run in sequence: a U-Net that outlines the vessels, a detector that finds the stenosis, and a stage that measures it. Because each is trained independently, an error early in the chain propagates and is amplified, which is the central weakness this baseline exposes. Flow 2 is the core contribution: a single model, **GCT-Net**, that performs all three tasks together. One shared MaxViT encoder feeds three heads, two that outline the vessels and the stenosis and one that measures the stenosis by focusing only on the detected lesion. A stop-gradient prevents the measuring head from distorting the detection, and the three task losses are balanced automatically rather than by hand. Flow 3 extends GCT-Net into a multi-source model. Alongside the angiogram, it incorporates two inputs a cardiologist also consults, the ECG and the written procedural report, merging them through a lightweight fusion layer (FiLM). It adds four report-derived heads and wraps the measurements in reliable confidence intervals, broadening the output at a small cost in image accuracy. The results show the benefit of joint optimisation: Flow 2 raises the stenosis overlap score (Dice) from 0.465 to 0.610 and lowers the measurement errors from 18.9 pp and 5.7 mm to 11.73 pp and 4.51 mm in a single pass, while Flow 3 trades a little accuracy for broader clinical output. The model is also validated on a separate public dataset (ARCADE), reaching a Dice of 0.523 on 300 unseen images, and is deployed as a web application that converts raw DICOM archives into a structured QCA report.

Keywords: Coronary Artery Disease, Deep Learning, Multi-Task Learning, Multi-Source Fusion, MaxViT, FiLM, Quantitative Coronary Analysis, Conformal Prediction, External Validation.

Résumé

Les maladies cardiovasculaires sont la première cause de mortalité dans le monde, la maladie coronarienne en étant la forme la plus fréquente. Pour la diagnostiquer, les cliniciens utilisent l'angiographie coronarienne par rayons X, dont l'interprétation manuelle est lente et sujette à des désaccords entre experts. La plupart des outils d'apprentissage profond traitent les trois sous-tâches (délimiter les vaisseaux, repérer la sténose et mesurer sa sévérité par l'analyse coronarienne quantitative, QCA) avec des modèles distincts entraînés séparément. Pour l'angiographie par rayons X, aucun réseau unique ne produit encore les trois sorties à la fois, ne fusionne des données non-imagerie et ne rapporte sa confiance.

Cette thèse conçoit et compare trois architectures d'ambition croissante, appelées *flux*, sur une cohorte de 200 patients. Le Flux 1 est une référence où trois modèles distincts s'enchaînent : un U-Net qui délimite les vaisseaux, un détecteur qui repère la sténose et une étape qui la mesure. Chaque modèle étant entraîné séparément, une erreur en début de chaîne se propage et s'amplifie, faiblesse centrale révélée par cette référence.

Le Flux 2 est la contribution principale : un modèle unique, **GCT-Net**, qui réalise les trois tâches ensemble. Un encodeur MaxViT partagé alimente trois têtes, deux qui délimitent les vaisseaux et la sténose et une qui mesure la sténose en se concentrant sur la lésion détectée. Un stop-gradient empêche la tête de mesure de déformer la détection, et les pertes des trois tâches sont équilibrées automatiquement plutôt qu'à la main.

Le Flux 3 étend GCT-Net en un modèle multi-source. Outre l'angiogramme, il intègre deux entrées que le cardiologue consulte aussi, l'ECG et le compte-rendu, fusionnés via une couche légère de fusion (FiLM). Il ajoute quatre têtes dérivées du compte-rendu et entoure les mesures d'intervalles de confiance fiables, élargissant les sorties au prix d'une légère baisse de précision en imagerie.

Les résultats montrent l'apport de l'optimisation conjointe : le Flux 2 fait passer le Dice sténose de 0,465 à 0,610 et réduit les erreurs %DS et longueur de 18,9 pp et 5,7 mm à 11,73 pp et 4,51 mm en une seule passe, tandis que le Flux 3 échange de la précision contre une sortie clinique plus large. Le modèle est aussi validé sur un jeu public indépendant (ARCADE), atteignant un Dice de 0,523 sur 300 images inédites, et il est déployé dans une application web qui convertit des archives DICOM brutes en un rapport QCA structuré.
Mots clés : Maladie Coronarienne, Deep Learning, Apprentissage Multi-Tâches, Fusion Multimodale, MaxViT, FiLM, QCA, Prédiction Conforme, Validation Externe.

Contents

List of Figures **v**

List of Tables **vii**

List of Abbreviations **ix**

Introduction **1**

1 Preliminaries **5**

1.1 Introduction 5
1.2 The Heart: Anatomy and Coronary Circulation 5
1.3 Cardiovascular Disease 6
1.3.1 Overview and Epidemiology 6
1.3.2 Coronary Artery Disease (CAD) 6
1.3.3 Diagnostic Approaches and Medical Reporting 7
1.3.4 Coronary Intervention and Treatment Strategy 9
1.4 Medical Imaging Modalities 10
1.4.1 Electrocardiography (ECG) 10
1.4.2 Computed Tomography Angiography (CCTA) 11
1.4.3 X-ray Coronary Angiography (ICA) 11
1.5 Digital Image Processing 12
1.5.1 Image Preprocessing 12
1.5.2 Image Segmentation 14
1.6 Deep Learning 15
1.6.1 Optimization Algorithms 15
1.6.2 Convolutional Neural Networks (CNNs) 16
1.6.3 Transformer Architectures 17
1.6.4 Multi-Task Learning 18
1.6.5 Loss Functions 20
1.6.6 Transfer Learning 24
1.6.7 Data Augmentation 26
1.6.8 Mixed-Precision Training 26

1.7 Conclusion . . . 27

2 Deep Learning for Cardiovascular Disease Detection: State of the Art 28
2.1 Introduction . . . 28
2.2 Datasets and Evaluation Metrics . . . 29
2.2.1 Datasets . . . 29
2.2.2 Evaluation Metrics . . . 29
2.3 Literature Review . . . 32
2.3.1 Vessel Segmentation . . . 32
2.3.2 Lesion Detection . . . 33
2.3.3 Lesion and Stenosis Quantification . . . 35
2.3.4 Multi-Task Approaches in Coronary and Vessel Analysis . . . 36
2.4 Comparative Analysis . . . 38
2.5 Conclusion . . . 40

3 Proposed Approach 41
3.1 Introduction . . . 41
3.2 Dataset Description . . . 43
3.3 Preprocessing and Ground-Truth Creation . . . 45
3.3.1 Image Preprocessing . . . 45
3.3.2 Frame Sorting . . . 45
3.3.3 Segmentation Ground-Truth Creation . . . 46
3.3.4 Non-Imaging Feature Extraction . . . 47
3.3.5 Per-Patient Dataset Organisation . . . 49
3.3.6 Multi-Source Ground Truth . . . 50
3.4 Methodology: The Three Architectural Flows . . . 51
3.5 Flow 1 — Sequential Pipeline . . . 52
3.6 Flow 2 — Single-Source / Multi-Task Architecture: GCT-Net . . . 66
3.6.1 Design Rationale . . . 66
3.6.2 Novel Contributions . . . 68
3.6.3 Automatic Best-Frame Selection . . . 71
3.6.4 Shared MaxViT Encoder . . . 74
3.6.5 Vessel and Stenosis Decoders . . . 75
3.6.6 Lesion-Aware QCA Head . . . 76
3.6.7 Stop-Gradient between the Stenosis Decoder and the QCA Head . . . 77
3.7 Multi-Task Loss with Homoscedastic Uncertainty Weighting . . . 78
3.8 Training Strategy . . . 80
3.8.1 Stratified Patient-Level Split and the Role of the Held-Out Set . . . 80
3.8.2 Data Augmentation Pipeline . . . 81
3.8.3 Optimiser and Discriminative Learning Rates . . . 81

3.8.4 Three-Stage Training Curriculum 82
3.9 Inference Pipeline 83
3.10 Experimental Results 84
3.10.1 Training Dynamics 84
3.10.2 Segmentation Results 85
3.10.3 QCA Quantification Results 87
3.10.4 Qualitative Results 87
3.10.5 Limitations of the Single-Source Model 90
3.11 Flow 3 — Multi-Source / Multi-Task Extension: GCT-Net 91
3.11.1 Heterogeneous Inputs 91
3.11.2 Modality-Specific Encoders and FiLM Fusion 92
3.11.3 Novel Contributions 93
3.11.4 Auxiliary Report-Derived Heads 95
3.11.5 Seven-Task Loss, Annealable Coupling, clDice and Conformal Calibration 96
3.11.6 Four-Stage Curriculum and Inference 99
3.11.7 Results 99
3.11.8 Limitations of the Multi-Source Extension 101
3.12 Comparative Analysis of the Three Flows 104
3.12.1 Qualitative Architectural Comparison 104
3.12.2 Quantitative Head-to-Head 105
3.12.3 The Central Trade-off: Imaging Accuracy versus Clinical Breadth . 107
3.13 External Validation on the ARCADE Dataset 108
3.13.1 Dataset and Protocol 108
3.13.2 Pre-Adaptation Cross-Domain Evaluation 110
3.13.3 Fine-Tuning on ARCADE 110
3.13.4 Re-Test after Fine-Tuning 110
3.13.5 Discussion 112
3.14 Web Application Deployment 112
3.15 Discussion 115
3.15.1 Effects of the Multi-Task Design 115
3.15.2 Effect of the Stop-Gradient and the Annealable Coupling 115
3.15.3 Effect of the Centerline-Dice Loss (Flow 3) 116
3.15.4 Multi-Source Supervision under Partial Labelling 116
3.15.5 Calibration via Conformal Prediction (Flow 3) 116
3.15.6 Limitations 116
3.16 Conclusion 118

General Conclusion **120**

Appendix A: Annex: Deposit License Certificate 127

List of Figures

1.1 Heart anatomy and coronary arteries 6
1.2 QCA measurements on a stenosed segment 8
1.3 Example cardiology medical report 9
1.4 12-lead ECG of anterior STEMI 11
1.5 Effect of CLAHE on an angiography frame 13
1.6 Effect of Gaussian denoising 13
1.7 The three categories of image segmentation 14
1.8 ResNet residual block architecture 16
1.9 The U-Net architecture 17

3.1 Chosen frame for a representative patient 44
3.2 Annotated frame with QCA measurements 44
3.3 Binary vessel mask creation 46
3.4 Binary stenosis mask creation 47
3.5 Per-patient dataset folder structure 50
3.6 Flow 1 sequential pipeline architecture 53
3.7 Vessel segmentation Attempt 1 54
3.8 Two-pass background-suppression scheme (Attempt 2) 56
3.9 Vessel segmentation Attempt 3 (final model) 58
3.10 Stenosis localisation Attempt 1 59
3.11 Stenosis localisation Attempt 2 (vessel-gated) 60
3.12 Stenosis localisation Attempt 3 (final model) 62
3.13 Example QCA quantification output 64
3.14 Flow 1 training dynamics 65
3.15 The cascading-error limitation of Flow 1 66
3.16 Overview of GCT-Net (Flow 2) 68
3.17 Contribution 1: lesion-aware ROI pooling 69
3.18 Contribution 2: the stop-gradient 70
3.19 Contribution 3: shared encoder, self-balancing loss 72
3.20 Automatic best-frame selection pipeline 72
3.21 The three blocks of a MaxViT-tiny stage 75

3.22 The five hierarchical feature maps of MaxViT-tiny 76
3.23 Homoscedastic uncertainty weighting of the task losses 79
3.24 The three-stage training curriculum (Flow 2) 83
3.25 Training dynamics of GCT-Net (Flow 2) 86
3.26 QCA regression error of GCT-Net (Flow 2) 87
3.27 Qualitative validation results of GCT-Net (Flow 2) 89
3.28 Limitation 1: asymmetric coupling 90
3.29 Limitation 2: brittle curriculum 90
3.30 Limitation 3: non-modular design 91
3.31 Architecture of GCT-Net (Flow 3) 92
3.32 Contribution 1: FiLM fusion 94
3.33 Contribution 2: multi-source supervision 94
3.34 Contribution 3: reporting, balanced loss, robustness 95
3.35 Training dynamics of GCT-Net (Flow 3) 100
3.36 Auxiliary-head learning and stenosis–QCA coupling 102
3.37 Limitation 1: negative transfer 102
3.38 Limitation 2: sparse, long-tailed report labels 103
3.39 Limitation 3: partially observed modalities 103
3.40 Comparative analysis of the three flows 106
3.41 Relative change of the unified flows vs. Flow 1 107
3.42 The two-stage ARCADE external-validation protocol 109
3.43 Fine-tuning dynamics on the ARCADE training split 111
3.44 Fine-tuned performance on the ARCADE test split 111
3.45 External validation on ARCADE 112
3.46 Web application: upload and pipeline overview 114
3.47 Web application: segmentation outputs 114
3.48 Web application: QCA quantification with conformal intervals 115
3.49 Empirical coverage of the split-conformal QCA predictor 117

List of Tables

2.1 Publicly available datasets for CVD detection using deep learning. 29
2.2 Comparative Analysis of Deep Learning Methods for Vessel Segmentation. SN: Sensitivity; SP: Specificity; XRA: X-ray angiography; CTA: CT angiography; ICA: non-invasive coronary angiography. 33
2.3 Comparative Analysis of Deep Learning Methods for Lesion Detection. . . 35
2.4 Comparative Analysis of Deep Learning Methods for Lesion/Stenosis Quantification. 36
2.5 Multi-task and joint quantification approaches in coronary and vascular analysis. 38
2.6 Comparison of Deep Learning Methods for Cardiovascular Disease Detection. 39

3.1 The twelve ECG features extracted by NeuroKit2 from each 12-lead tracing, together with the extraction-confidence score. Binary flags are encoded as 0/1; the confidence score lies in $[0, 1]$. 48
3.2 The structured tokens extracted from each free-text procedural report by the large-language-model parser. Categorical fields are later one-hot encoded, numeric fields normalised, and the treated-vessel field treated as a multi-label over the major coronary arteries. 49
3.3 Multi-source ground-truth signals used to train GCT-Net. 50
3.4 Quantitative results of the Flow 1 sequential pipeline. The vessel stage carries no withheld manual masks and is reported by its converged training loss; the stenosis and QCA stages are evaluated on $n = 20$ held-out patients at the tuned operating threshold $\tau = 0.15$. The stenosis Dice differs between Phase 2 (0.345) and the Phase 3 segmentation head (0.465) because they are two separately trained models. 65
3.5 Three-stage training schedule used for GCT-Net (Flow 2). Batch size is 4 on a single Tesla T4 GPU; η_{head} denotes the head learning rate, with the encoder at half this rate. A hard stop-gradient is applied at the stenosis-to-QCA junction throughout all three stages. 82

3.6 Best validation metrics achieved during each training stage of GCT-Net (Flow 2) on fold 0. Vessel Dice is undefined (no withheld manual vessel masks). 85
3.7 Segmentation results of GCT-Net (Flow 2) on the validation split (fold 0). Stenosis Dice is reported as mean ± standard deviation under four-flip TTA and the adaptive threshold. 86
3.8 QCA quantification results of GCT-Net (Flow 2) on the validation split (fold 0, $n = 21$ patients with valid DICOM QCA tags), reported as mean absolute error of the de-normalised predictions. 87
3.9 Results of GCT-Net (Flow 3) on the held-out validation fold (fold 0). Conformal coverage is at a nominal 90% target ($\alpha = 0.10$). 101
3.10 Qualitative comparison of the three flows along five architectural dimensions. The progression moves from many independent models toward a single unified network, from cascaded toward joint optimisation, and from a single imaging modality toward multi-source clinical input. 105
3.11 Comparative analysis of the three flows. Stenosis Dice for Flow 1 uses the Phase 3 multi-task model; the Flow 2/3 figures are the unified MaxViT-tiny multi-task models. "pp" denotes percentage points of diameter stenosis. Vessel Dice is unavailable across all flows because no manual vessel masks were withheld for validation. Best value in each imaging row is shown in bold. 106
3.12 External validation on the ARCADE coronary-stenosis test split ($n = 300$ images, operating threshold 0.50) after fine-tuning. The Oasis-trained model transfers poorly zero-shot (not clinically usable; see Section 3.13.2); fine-tuning on the 850-image ARCADE training split (best ival checkpoint, epoch 101) recovers the performance reported below. The vessel-gating multiplication is disabled for all ARCADE inference. 111

List of Abbreviations

AI	Artificial Intelligence
AMP	Automatic Mixed Precision
ANN	Artificial Neural Network
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
BMS	Bare-Metal Stent
CABG	Coronary Artery Bypass Grafting
CAD	Coronary Artery Disease
CCC	Concordance Correlation Coefficient (Lin's CCC)
CCTA	Coronary Computed Tomography Angiography
CLAHE	Contrast Limited Adaptive Histogram Equalization
clDice	Centerline Dice (topology-preserving Dice variant)
CNN	Convolutional Neural Network
CP	Conformal Prediction
CT	Computed Tomography
CTO	Chronic Total Occlusion
CVD	Cardiovascular Disease
DES	Drug-Eluting Stent
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DS	Diameter Stenosis
ECG	Electrocardiogram
ESC/EACTS	European Society of Cardiology / European Association for Cardio-Thoracic Surgery
FFR	Fractional Flow Reserve
FN	False Negative
FP	False Positive
FT	Focal Tversky (loss)
GAN	Generative Adversarial Network
GCT-Net	Global Coronary Transformer Network (proposed model)

HD	Hausdorff Distance
ICA	non-invasiveCoronary Angiography
IoU	Intersection over Union
LAD	Left Anterior Descending artery
LCx	Left Circumflex artery
LM	Left Main coronary artery
LoA	Limits of Agreement (Bland–Altman)
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MaxViT	Multi-axis Vision Transformer
MBConv	Mobile Inverted Bottleneck Convolution
ML	Machine Learning
MLD	Minimal Lumen Diameter
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MTL	Multi-Task Learning
NSTEMI	Non-ST-Elevation Myocardial Infarction
PCI	Percutaneous Coronary Intervention
PI	Prediction Interval (conformal)
QCA	Quantitative Coronary Analysis
R-CNN	Region-based Convolutional Neural Network
RCA	Right Coronary Artery
ReLU	Rectified Linear Unit
ResNet	Residual Network
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RNN	Recurrent Neural Network
RVD	Reference Vessel Diameter
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Index Measure
SSL	Self-Supervised Learning
STEMI	ST-Elevation Myocardial Infarction
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TTA	Test-Time Augmentation
ViT	Vision Transformer
WHO	World Health Organization

%DS	Percent Diameter Stenosis
YOLO	You Only Look Once

Introduction

Context and Motivation

Cardiovascular diseases (CVDs) are the leading cause of global mortality, claiming more lives annually than any other category of disease. According to the World Health Organization, CVDs are responsible for approximately 19.8 million deaths each year, representing 32% of all global deaths [1]. Coronary artery disease (CAD) is the most prevalent form, characterised by the progressive buildup of atherosclerotic plaques that narrow the coronary arteries and restrict myocardial perfusion.

The diagnosis of CAD relies on a multi-task workflow, ranging from non-invasive screening with Coronary Computed Tomography Angiography (CCTA) to the gold standard of non-invasive Coronary Angiography (ICA). Stenosis quantification remains, however, problematic. In routine practice the assessment is often performed visually, which is subjective and prone to inter-observer variability. Objective methods such as Quantitative Coronary Analysis (QCA) provide accurate measurements but are labour-intensive, creating a bottleneck in time-critical clinical workflows [8].

Coronary stents play a central role in the interventional treatment of CAD. A stent is a small, mesh-like metallic tube inserted into a narrowed coronary artery during Percutaneous Coronary Intervention (PCI) to mechanically restore blood flow. Modern Drug-Eluting Stents (DES) are coated with antiproliferative agents that prevent restenosis. Stent deployment must be precisely planned: the stent diameter must match the reference vessel diameter to within a fraction of a millimetre, and the stent length must fully cover the lesion plus a healthy margin on both sides. Undersizing leads to malapposition and increased thrombosis risk, whereas oversizing risks vessel dissection or rupture [9].

This sizing requirement is particularly demanding during PCI, where the interventional cardiologist must take rapid, high-stakes decisions on stent number, diameter and length. Determining whether a long lesion requires a single stent or multiple overlapping stents and identifying the precise landing zones in major arteries such as the Left Anterior Descending (LAD) calls for a level of precision that exceeds what visual estimation alone can reliably provide [9].

Deep learning [2] has become the dominant tool for medical image analysis. Hybrid architectures that combine the local feature extraction of Convolutional Neural Networks

(CNNs) [30] with the long-range modelling capacity of Transformers [31] are particularly well suited to coronary imaging, where vessel boundaries are local but the topology of the coronary tree is global. They offer a route towards automated, objective decision support that can quantify clinical metrics in real time [8].

Problem Statement

Despite advances in medical imaging, accurate and timely diagnosis of coronary artery disease (CAD) remains a complex challenge. Clinical workflows in high-volume centres such as Oasis Clinic exhibit several limitations:

1. **Subjectivity and inter-observer variability.** Visual assessment of stenosis severity is the current norm in many centres but suffers from substantial variability between clinicians, which can lead to suboptimal decisions on revascularisation.

2. **Quantitative precision.** Stent selection requires millimetre-level measurements of vessel diameter and lesion length. Manual QCA provides this accuracy but is too time-consuming for routine acute cases.

3. **2D projection limitations.** X-ray angiography projects a three-dimensional vascular tree onto a two-dimensional plane, leading to overlap and foreshortening; distinguishing the relevant artery from background structures and overlapping vessels remains a difficult task for standard algorithms.

4. **Workflow bottlenecks.** In acute settings such as ST-Elevation Myocardial Infarction (STEMI), every minute of ischaemia causes irreversible myocardial damage; the absence of automated, real-time decision support introduces avoidable delays in the catheterisation laboratory.

5. **Gap in end-to-end ICA analysis.** Multi-task formulations have emerged for coronary CT angiography and for 3D vascular volumes, but on X-ray non-invasive coronary angiography (ICA) the dominant systems still rely on sequential pipelines in which stenosis quantification is derived from post-hoc geometric algorithms applied to the segmentation output rather than from a jointly learned regression head. To the best of our knowledge, no prior work on ICA jointly optimises vessel segmentation, stenosis localisation, and continuous QCA regression within a single end-to-end network.

Objectives

The objective of this work is to investigate and compare three progressively ambitious deep-learning architectures for the automated analysis of X-ray non-invasive coronary an-

giography (ICA) — vessel segmentation, stenosis localisation, and Quantitative Coronary Analysis (QCA) — culminating in a unified multi-source, multi-task system. The specific aims are:

- *Establish a sequential baseline (Flow 1):* reproduce the dominant ICA-quantification paradigm by chaining three independently trained models — a U-Net vessel segmenter, a stenosis detector, and a QCA regression stage — and quantify the cascading-error cost of this design empirically.

- *Design a single-source multi-task architecture (Flow 2):* propose a single network in which a shared hybrid CNN-Transformer encoder (MaxViT-tiny [15, 27]) feeds two parallel U-Net decoders for vessel and stenosis segmentation, and a lesion-aware ROI pooling head that conditions QCA regression on the predicted stenosis map. A stop-gradient decouples the localisation map from the regression objective, and homoscedastic uncertainty weighting [16] balances the task losses automatically. The model is trained under partial supervision from heterogeneous label sources — manual masks, U-Net pseudo-labels, and DICOM-parsed QCA targets — using per-sample validity flags [39].

- *Extend to a multi-source, multi-task architecture (Flow 3):* fuse a tabular ECG branch and structured fields parsed from the procedural report into the same backbone through a FiLM bottleneck [34], add four report-derived supervision heads, and equip the QCA outputs with calibrated prediction intervals via a split-conformal predictor [56].

- *Evaluate and compare the three flows:* validate the system on a 200-patient cohort from Oasis Clinic (Ghardaïa, Algeria), reporting segmentation Dice, QCA Mean Absolute Error, correlation coefficients, Bland–Altman agreement, and conformal coverage.

- *Assess cross-domain generalisability:* externally evaluate the stenosis branch on the independent ARCADE benchmark following a two-stage protocol — a zero-shot transfer test of the Oasis-trained model to quantify the domain gap, followed by fine-tuning on the ARCADE training split and re-evaluation on its 300-image test set to measure how effectively the learned representation transfers to unseen acquisition conditions.

- *Deploy the system in a clinical application:* integrate the trained model into a web application that ingests raw DICOM archives, automatically selects the best frame, runs the multi-task inference, and generates a structured QCA report.

Thesis Organisation

This thesis is structured into three chapters:

Chapter 1: Preliminaries establishes the clinical and theoretical foundations. It covers coronary anatomy and pathophysiology, interventional treatment including stenting, diagnostic standards including medical reporting and Quantitative Coronary Analysis, and the three modalities employed in this work (ECG, CCTA, X-ray ICA) with the DICOM standard. It then develops the relevant deep learning foundations: CNN and Transformer architectures, the MaxViT hybrid encoder, multi-task and multi-source learning under partial supervision, FiLM fusion, homoscedastic uncertainty weighting, the task-specific loss functions used in the proposed system, conformal prediction, and the supporting training machinery (transfer learning, knowledge distillation, augmentation, and mixed-precision training).

Chapter 2: State of the Art surveys deep learning methods for cardiovascular disease detection across vessel segmentation, lesion detection, and stenosis quantification, analysing the architectures, datasets, results, and limitations for each. A dedicated subsection reviews multi-task and joint-quantification approaches on both CCTA and ICA. The chapter identifies the gap motivating this work: on X-ray ICA, no existing system jointly optimises all three tasks within a single end-to-end network with continuous learned QCA regression, non-imaging modality fusion, or calibrated uncertainty bounds.

Chapter 3: Proposed Approach and Experimental Results presents three progressively ambitious architectures on a 200-patient cohort from Oasis Clinic (Ghardaïa, Algeria). Flow 1 chains three independently trained models as a sequential baseline, exposing the cascading-error limitation. Flow 2 proposes GCT-Net, a single-source multi-task model whose shared MaxViT-tiny encoder feeds two segmentation decoders and a lesion-aware ROI pooling head for QCA regression, with stop-gradient decoupling and uncertainty weighting under partial supervision. Flow 3 extends GCT-Net to a multi-source architecture fusing a tabular ECG branch and procedural-report fields via FiLM, adding four report-derived heads and a split-conformal predictor for calibrated QCA intervals. The chapter concludes with a comparative analysis of the three flows, external validation on the ARCADE benchmark, and deployment in a clinical web application.

The thesis closes with a synthesis of contributions and a discussion of future directions for improving generalisability and clinical integration.

Chapter 1

Preliminaries

1.1 Introduction

This chapter establishes the theoretical and clinical foundation for the research presented in this thesis. We begin by providing a brief overview of heart anatomy, followed by a clinical overview of cardiovascular disease, specifically focusing on coronary pathophysiology, interventional treatments such as stenting, and current diagnostic standards including medical reporting and Quantitative Coronary Analysis (QCA). We then introduce the principal diagnostic modalities employed Electrocardiography (ECG), Coronary CT Angiography (CCTA), and X-ray Invasive Coronary Angiography (ICA) — together with the DICOM standard and the digital image processing techniques used in the preprocessing pipeline. Finally, we provide a comprehensive examination of the machine learning and deep learning paradigms underpinning the proposed solution, detailing the mathematical formulations of Convolutional Neural Networks (CNNs), Transformer architectures, multi-task learning, and the loss functions, uncertainty estimation, and training techniques.

1.2 The Heart: Anatomy and Coronary Circulation

The heart is a muscular organ located in the mediastinum, responsible for pumping oxygenated blood to the entire body. It is composed of four chambers: two atria (right and left) that receive blood, and two ventricles (right and left) that eject blood into the pulmonary and systemic circulations, respectively [29].

The myocardium (heart muscle) itself is supplied with oxygenated blood by the *coronary arteries*, which arise from the root of the aorta. The three principal coronary vessels are the Left Anterior Descending (LAD) artery, which supplies the anterior wall of the left ventricle; the Left Circumflex (LCx) artery, which supplies the lateral wall; and the Right Coronary Artery (RCA), which supplies the inferior wall and the right ventricle.

Any narrowing (stenosis) of these vessels reduces myocardial perfusion and can precipitate ischaemia or infarction.

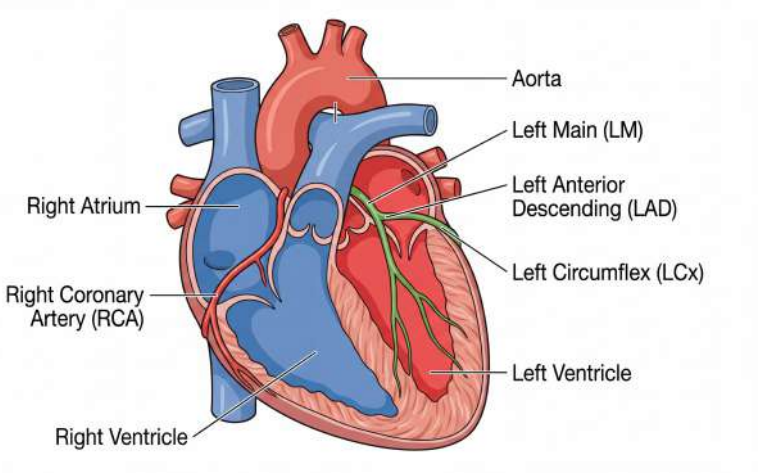


Figure 1.1: Schematic of heart anatomy showing the four chambers (right and left atria and ventricles) and the principal coronary arteries (Left Main, Left Anterior Descending, Left Circumflex, and Right Coronary Artery) arising from the root of the aorta. Adapted from [29].

1.3 Cardiovascular Disease

1.3.1 Overview and Epidemiology

Cardiovascular disease (CVD) encompasses a group of disorders affecting the heart and blood vessels, including coronary artery disease (CAD), cerebrovascular disease, and peripheral arterial disease. According to the World Health Organization (WHO), CVDs are the leading cause of death globally, responsible for an estimated 19.8 million deaths annually, representing approximately 32% of all global deaths. The primary driver is atherosclerosis, a chronic inflammatory process characterised by the deposition of lipids, cholesterol, and calcium within the arterial intima.

1.3.2 Coronary Artery Disease (CAD)

Coronary artery disease occurs when the coronary arteriesthe vessels responsible for supplying oxygenated blood to the myocardiumbecome narrowed (stenotic) or blocked by atherosclerotic plaques.

Stenosis Classification

The degree of arterial narrowing is expressed as a percentage of vessel diameter reduction. A reduction below 25% is termed *minimal* and generally requires no specific intervention

beyond risk-factor management. Between 25% and 49% the stenosis is classified as *mild* and is typically managed with medical therapy alone. Between 50% and 69% the stenosis is considered *moderate*; moderate lesions often require functional testing such as Fractional Flow Reserve (FFR) to determine whether blood flow is significantly impeded. Reductions of 70–99% are classified as *severe* and typically warrant intervention. A complete (100%) blockage is termed a *total occlusion*; if it persists for more than three months it is referred to as a Chronic Total Occlusion (CTO).

1.3.3 Diagnostic Approaches and Medical Reporting

The diagnosis of cardiovascular disease involves a multi-faceted approach ranging from clinical assessment to advanced imaging.

Quantitative Coronary Analysis (QCA)

In both clinical practice and automated diagnosis research, quantifying lesion severity is essential. Quantitative Coronary Analysis (QCA) [8] provides the objective reference metrics:

- *Minimal Lumen Diameter (MLD):* the diameter of the artery at the point of maximum narrowing (in mm).
- *Reference Vessel Diameter (RVD):* the estimated diameter of the vessel at the lesion site if it were healthy, usually interpolated from the proximal and distal healthy segments.
- *Percent Diameter Stenosis (%DS):* the standard clinical severity metric, which expresses the rate of vessel narrowing as a percentage of the reference diameter. It is defined as
$$\%DS = \left(1 - \frac{\text{MLD}}{\text{RVD}}\right) 100, \tag{1.1}$$
and can be interpreted directly as the fraction of the healthy lumen diameter that has been lost at the lesion site (e.g., $\%DS = 70$ indicates that the lumen has been reduced to 30% of its reference diameter).
- *Lesion Length:* the distance along the vessel centreline from the proximal to the distal shoulder of the lesion.

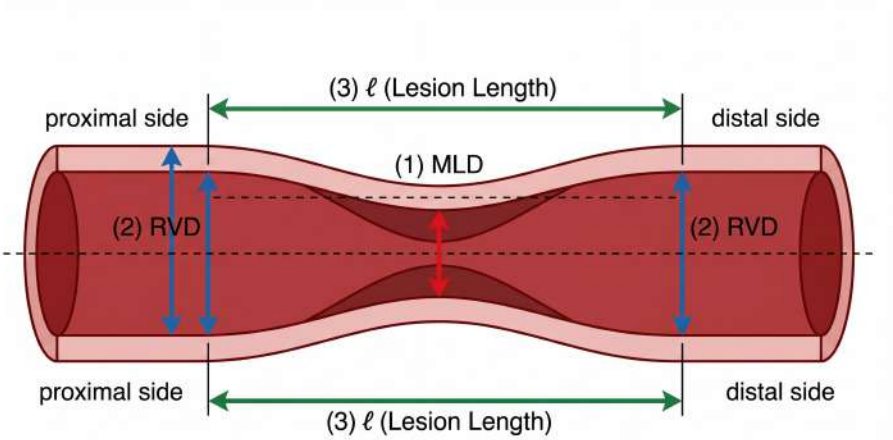
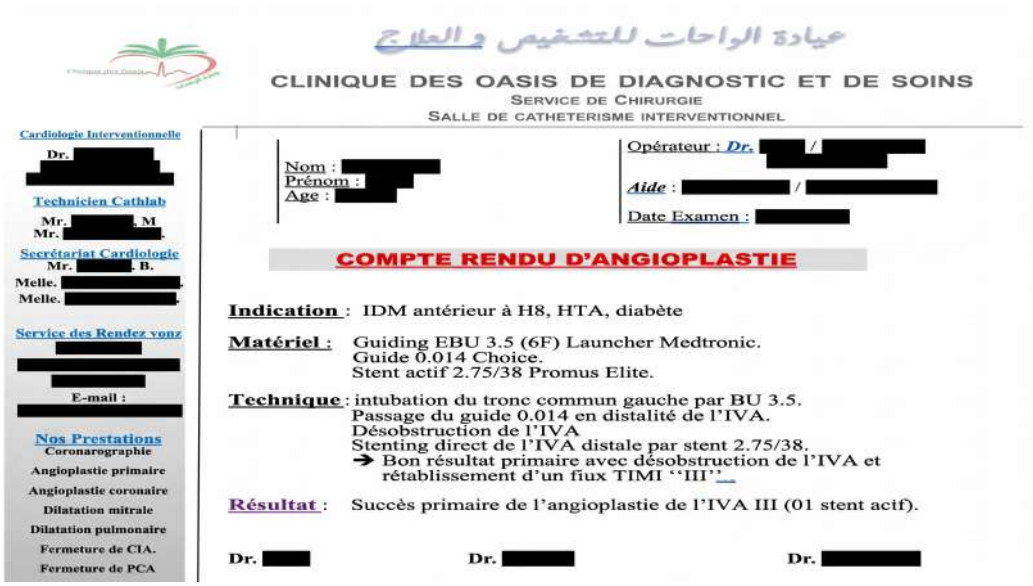


Figure 1.2: Illustrative example of Quantitative Coronary Analysis (QCA) measurements on a stenosed coronary segment. The Minimal Lumen Diameter (MLD) is measured at the point of maximum narrowing; the Reference Vessel Diameter (RVD) is interpolated from the adjacent healthy segments (proximal and distal); the Lesion Length (ℓ) is measured along the vessel centreline between the proximal and distal shoulders of the lesion. The Percent Diameter Stenosis is derived from these quantities through Equation (1.1).

The Cardiology Medical Report. The medical report (*compte-rendu*) is the official record of a diagnostic or interventional procedure. A standard cardiology report typically includes:

1. *Indication:* the clinical reason for the procedure (e.g., typical angina, positive stress test, STEMI).
2. *Haemodynamics:* non-invasive pressure measurements (e.g., aortic pressure 120/80 mmHg).
3. *Angiographic findings (per vessel):*
   - *Left Main (LM):* evaluation of the trunk.
   - *Left Anterior Descending (LAD):* description of lesions in proximal, mid, or distal segments.
   - *Left Circumflex (LCx):* evaluation of the lateral wall supply.
   - *Right Coronary Artery (RCA):* evaluation of the inferior wall supply.
4. *Conclusion:* a summary of significant findings and the final therapeutic decision (medical therapy, PCI, or CABG surgery).


عيادة الواحات للتشخيص و العلاج

CLINIQUE DES OASIS DE DIAGNOSTIC ET DE SOINS

SERVICE DE CHIRURGIE

SALLE DE CATHETERISME INTERVENTIONNEL

Cardiologie Interventionnelle
Dr. [redacted]

Technicien Cathlab
Mr. [redacted]. M
Mr. [redacted].

Secrétariat Cardiologie
Mr. [redacted]. B.
Melle. [redacted].
Melle. [redacted].

Service des Rendez vonz
[redacted]

E-mail : [redacted]

Nos Prestations
Coronarographie
Angioplastie primaire
Angioplastie coronaire
Dilatation mitrale
Dilatation pulmonaire
Fermeture de CIA.
Fermeture de PCA

Nom : [redacted]
Prénom :
Age :

Opérateur : Dr. [redacted] / [redacted]
Aide : [redacted] / [redacted]
Date Examen : [redacted]

COMPTE RENDU D'ANGIOPLASTIE

Indication : IDM antérieur à H8, HTA, diabète

Matériel : Guiding EBU 3.5 (6F) Launcher Medtronic.
Guide 0.014 Choice.
Stent actif 2.75/38 Promus Elite.

Technique : intubation du tronc commun gauche par BU 3.5.
Passage du guide 0.014 en distalité de l'IVA.
Désobstruction de l'IVA
Stenting direct de l'IVA distale par stent 2.75/38.
➔ Bon résultat primaire avec désobstruction de l'IVA et rétablissement d'un fiux TIMI ''III''.

Résultat : Succès primaire de l'angioplastie de l'IVA III (01 stent actif).

Dr. [redacted] Dr. [redacted] Dr. [redacted]

Figure 1.3: Example of an anonymised cardiology medical report from the Oasis Clinic catheterisation laboratory, illustrating the standard structure: indication, haemodynamics, per-vessel angiographic findings (LM, LAD, LCx, RCA), and conclusion with the therapeutic decision. All patient-identifying information has been removed.

1.3.4 Coronary Intervention and Treatment Strategy

Percutaneous Coronary Intervention (PCI), commonly known as angioplasty, is the non-surgical procedure used to treat stenotic coronary arteries. The decision to proceed with stenting depends on lesion severity and clinical context, so not every stenosis is an indication for PCI. The choice between medical therapy, PCI and Coronary Artery Bypass Grafting (CABG) is governed by established clinical guidelines such as ESC/EACTS [9]:

- *Medical therapy:* preferred for stable CAD with non-significant stenosis.
- *PCI (stenting):* the standard approach for significant single- or double-vessel disease.
- *CABG (surgery):* typically preferred for complex multi-vessel disease or Left Main disease, often quantified using the SYNTAX Score[1].

When PCI is indicated, a stent — a small, expandable wire mesh tube — is deployed to keep the artery open [9].

- *Drug-Eluting Stents (DES):* modern stents are coated with medication (e.g., everolimus, zotarolimus) that is slowly released to prevent cell proliferation and restenosis. They have largely replaced older Bare-Metal Stents (BMS).

[1]The SYNTAX Score is an angiography-based anatomical complexity score used to grade the extent and complexity of coronary artery disease and to guide the choice between PCI and CABG.

- *Stent sizing:* selecting the correct stent size is critical for procedure success.
  - *Diameter* typically ranges from 2.0 mm to 5.0 mm, matching the calibre of human coronary arteries; small distal branches such as the distal LAD or LCx are often ∼2.0–2.5 mm, whereas proximal segments and the Left Main can reach 4.0–5.0 mm. Undersizing leads to stent malapposition (with elevated thrombosis risk), while oversizing risks vessel rupture [9].
  - *Length* typically ranges from 8 mm to 48 mm. Short stents (8–15 mm) are used for focal lesions, while longer stents (38–48 mm) cover diffuse disease. The stent must cover the entire lesion plus a healthy margin on both proximal and distal ends (the *landing zone*); when a single stent is insufficient, two overlapping stents are deployed.

1.4 Medical Imaging Modalities

Medical imaging constitutes the cornerstone of coronary artery disease diagnosis and follow-up. The following modalities are most commonly employed in clinical practice and form the basis of the datasets used in this work.

1.4.1 Electrocardiography (ECG)

Electrocardiography is a non-invasive technique that records the electrical activity of the heart over time [42]. While it does not directly visualise the coronary arteries, it is crucial for the initial localisation of ischaemia:

- *STEMI (ST-Elevation Myocardial Infarction)* indicates full-thickness (transmural) damage, usually caused by acute total occlusion; the leads showing ST elevation pinpoint the culprit artery (e.g., elevation in leads V1–V4 suggests LAD occlusion).
- *NSTEMI (Non-ST-Elevation Myocardial Infarction)* suggests partial blockage or subendocardial ischaemia, often associated with severe stenosis but not necessarily with total occlusion.

The resting ECG also yields a set of structured numeric measurements — heart rate, the RR, PR, QRS, QT and corrected QT (QTc) intervals, the P/R/T-wave amplitudes, ST deviation, and heart-rate variability — which can be digitised into a compact tabular feature vector. Such a vector provides a non-visual, complementary view of the patient's Heart state and is used as an auxiliary input modality in the multi-source model of Chapter 3.

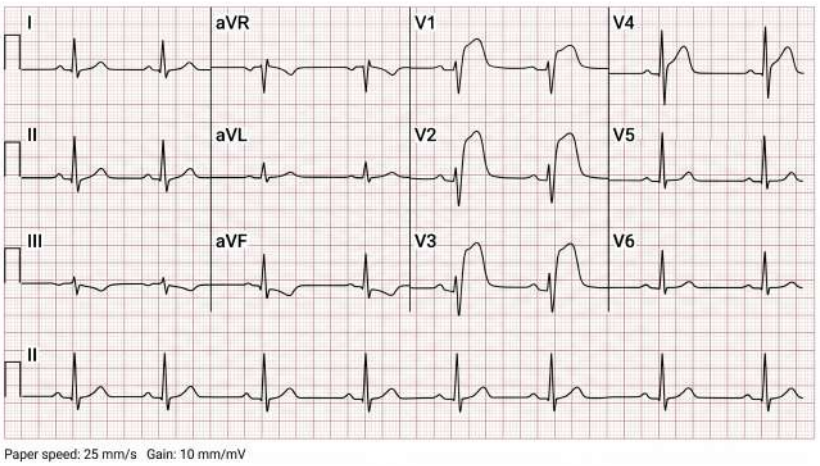


Figure 1.4: Representative 12-lead ECG showing an anterior STEMI pattern, with characteristic ST-segment elevation in the precordial leads (V1–V4) consistent with acute occlusion of the Left Anterior Descending artery.

1.4.2 Computed Tomography Angiography (CCTA)

Coronary Computed Tomography Angiography (CCTA) is a non-invasive modality that provides three-dimensional visualisation of the coronary tree [43].

- *Calcium scoring (Agatston score):* a non-contrast scan performed first to quantify calcified plaque burden. A score above 400 Agatston Units indicates a high plaque burden and may reduce diagnostic accuracy due to "blooming artefacts" (calcium obscuring the lumen) [43].
- *Plaque characterisation (Hounsfield Units, HU):* CCTA uses X-ray attenuation values to characterise plaque composition, with soft (lipid-rich, ≈30–70 HU) plaque being potentially unstable, fibrous plaque (≈70–130 HU) being more stable, and calcified plaque (>130 HU) being hardened.

1.4.3 X-ray Coronary Angiography (ICA)

non-invasive Coronary Angiography (ICA) remains the gold standard for guiding interventions. During the procedure, a radiopaque contrast agent is injected into the coronary arteries while X-ray fluoroscopy sequences are acquired. Since the coronary tree is a 3D structure projected onto a 2D plane, multiple camera angles are required to unmask overlapping vessels; common views include the Right Anterior Oblique (RAO) and Left Anterior Oblique (LAO) combined with cranial or caudal angulation.

The DICOM Standard

Medical images are stored and exchanged according to the Digital Imaging and Communications in Medicine (DICOM) standard. A DICOM file encapsulates pixel data and a

rich metadata header. The metadata fields most relevant to the present work are:

- *Patient and study information:* patient ID, acquisition date, institution name, and modality type.
- *Image geometry:* number of frames (for multi-frame cine sequences), rows, columns, and bits allocated per pixel.
- *Pixel spacing:* the physical distance (in mm) between the centres of two adjacent pixels in the detector plane. This calibration is essential for converting pixel-based measurements into real-world millimetres for stent sizing.
- *Photometric interpretation:* specifies whether pixel data is greyscale (MONOCHROME2) or colour (RGB).
- *Transfer syntax:* the compression scheme applied to the pixel data (e.g., JPEG 2000 lossless, used in the Oasis Clinic dataset).
- *Private tags:* vendor-specific fields written by the angiography console. In the Oasis Clinic dataset these encode the QCA measurements (reference vessel diameter, minimal lumen diameter and lesion length) used as regression targets in Chapter 3.

In this work, all DICOM files are parsed using the `pydicom` library[2] with a GDCM backend for JPEG 2000 decompression [12].

1.5 Digital Image Processing

Digital image processing encompasses techniques for manipulating and analysing digital images to extract clinically relevant information.

1.5.1 Image Preprocessing

Preprocessing is essential to improve image quality before applying deep learning models. The main preprocessing steps used in this work are:

- *Contrast Limited Adaptive Histogram Equalization (CLAHE)* is a standard technique for angiographic images [11]. Unlike global histogram equalisation, CLAHE divides the image into a regular grid of small tiles and applies local histogram equalisation to each tile, enhancing local contrast without amplifying noise in homogeneous regions. A clip-limit parameter (typically 2.0) prevents over-amplification of noise peaks.

[2] `pydicom` is an open-source Python package for reading, modifying, and writing DICOM files; project page: `https://pydicom.github.io/`.

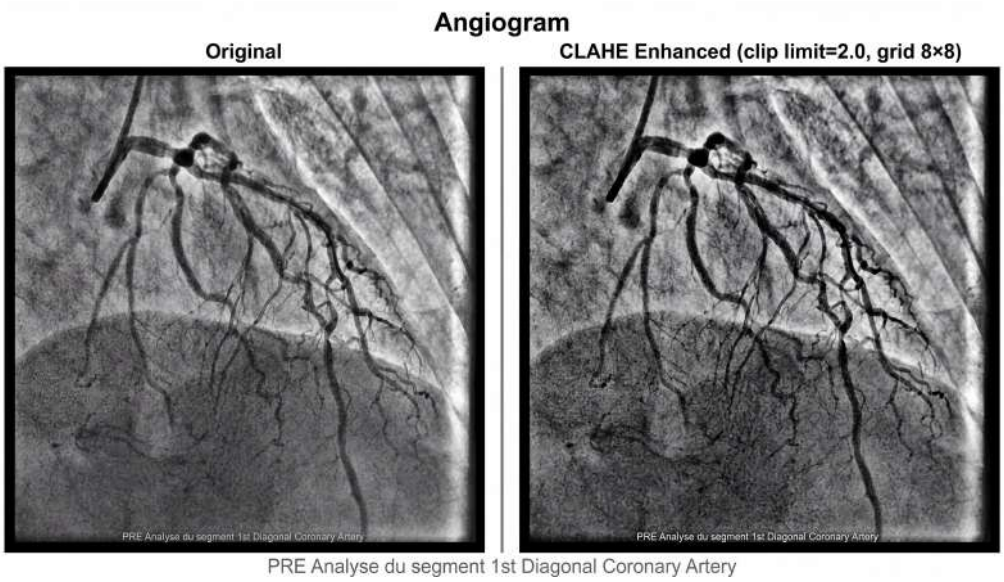


Figure 1.5: Effect of CLAHE on a coronary angiography frame. Left: original image. Right: CLAHE-enhanced image (clip limit = 2.0, tile grid 88). Local contrast within the coronary tree is increased without amplifying noise in homogeneous background regions.

- *Noise reduction.* Gaussian filtering suppresses Gaussian noise by convolving the image with a Gaussian kernel, while median filtering is preferred for impulsive (salt-and-pepper) noise as it preserves edge sharpness. In this work, a mild Gaussian blur with a 33 kernel is applied after CLAHE to reduce residual noise while preserving vessel edges.

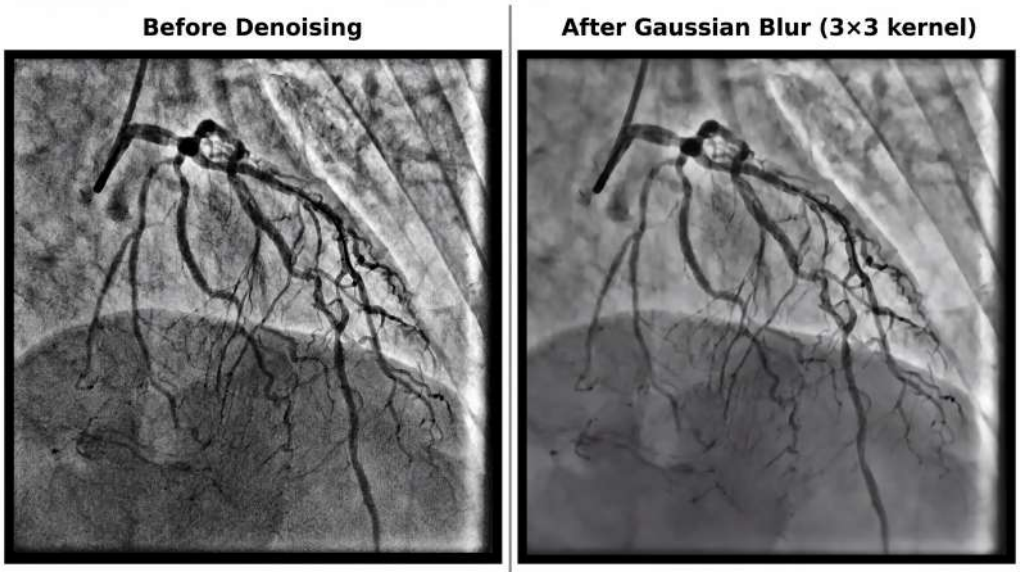


Figure 1.6: Effect of Gaussian denoising on a CLAHE-enhanced angiographic frame. Left: image before denoising. Right: image after a 33 Gaussian blur. Residual sensor noise is attenuated while vessel edges remain sharp.

1.5.2 Image Segmentation

Segmentation partitions an image into meaningful regions of interest. Before the rise of deep learning, segmentation methods were predominantly classical and could be grouped into three families: threshold-based methods, which classify pixels by their intensity; edge-based methods, which rely on intensity discontinuities to delineate region boundaries; and region-based methods, which grow or merge regions according to a homogeneity criterion. Two operations from these classical families remain useful in the present work as post-processing stages and are recalled below.

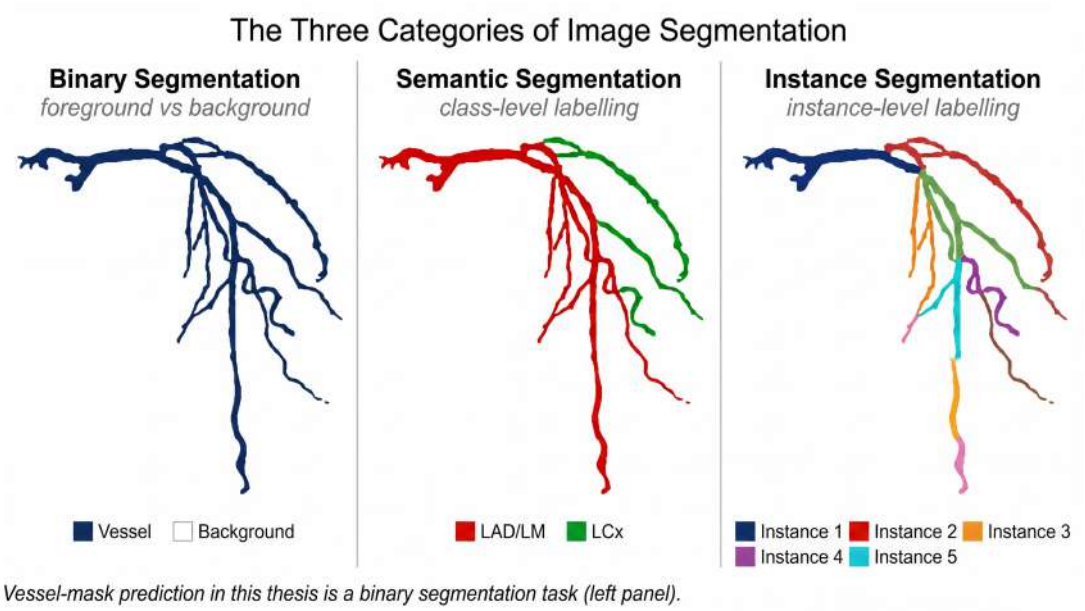


Figure 1.7: The three main categories of image segmentation. *Binary segmentation* (left) labels each pixel as foreground or background. *Semantic segmentation* (centre) assigns a class label to each pixel but does not distinguish between different instances of the same class. *Instance segmentation* (right) further separates individual instances of each class. Vessel-mask prediction in this thesis is a binary segmentation task.

Independently of the algorithmic family, the output of a segmentation method falls into one of three categories (Figure 1.7). *Binary segmentation* produces a two-class map separating an object of interest from the background. *Semantic segmentation* extends this to multiple classes by assigning a class label to every pixel, but it is not aware of object instances — all pixels belonging to the same class are labelled identically. *Instance segmentation* goes one step further by separating distinct instances of the same object class, so that two adjacent objects of the same category receive different labels. The vessel- and stenosis-prediction tasks addressed in this work are both binary segmentation problems.

Thresholding. The simplest and most widely used classical segmentation technique classifies pixels as foreground or background according to whether their intensity exceeds

a threshold value T. Although global thresholding is often insufficient for angiographic images because vessel intensity varies along the coronary tree, it remains a useful component when combined with local adaptive thresholding or with downstream morphological refinement.

Morphological operations. These operations are essential for vessel analysis. *Erosion* removes pixels on object boundaries and helps separate touching vessels, while *dilation* adds pixels to boundaries and fills small gaps in vessel contours. *Skeletonisation* iteratively erodes a vessel structure until a one-pixel-wide centreline remains, which is the standard input for downstream diameter and length measurement.

1.6 Deep Learning

Machine learning is the broader discipline that enables computers to learn statistical patterns from data and generalise these patterns to unseen examples [32]. *Deep learning* is a sub-field of machine learning in which models are based on multi-layered (deep) neural networks that automatically learn hierarchical representations of the input data [2]. Because the present work relies exclusively on deep learning, we limit our discussion in this chapter to the deep-learning concepts directly relevant to coronary artery analysis. The key components used in our work are described below.

1.6.1 Optimization Algorithms

The optimisation algorithm controls how the network parameters are updated from the loss gradient at each training step.

Stochastic Gradient Descent (SGD). The classical method computes the gradient of the loss function on a random mini-batch of samples and updates the parameters in the negative-gradient direction, scaled by a learning rate [2]. SGD is simple and well-understood but converges slowly when the loss surface is ill-conditioned.

Adam (Adaptive Moment Estimation). Adam computes adaptive per-parameter learning rates by combining first- and second-moment estimates of the gradients [2], and is the standard optimiser for medical image analysis. Its update rule is

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t, \tag{1.2}$$

where θ_t is the parameter vector at step t, η the learning rate, $\hat{m}_t$ the bias-corrected first moment (mean) of the gradients, $\hat{v}_t$ the bias-corrected second moment (uncentred variance) of the gradients, and ϵ a small constant (typically 10^{-8}) for numerical stability.

AdamW. AdamW [14] is a variant of Adam in which the L2 weight-decay term is decoupled from the gradient-based update and applied directly to the parameters. This decoupling restores the intended regularisation behaviour of weight decay for adaptive optimisers and improves generalisation; AdamW is the optimiser used to train the proposed models in Chapter 3.

1.6.2 Convolutional Neural Networks (CNNs)

CNNs are neural networks designed for grid-structured data such as images. They apply learnable convolutional filters that extract local spatial features, pooling layers that achieve spatial invariance, and non-linear activations [2].

- *ResNet* [30] introduces residual connections ($\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$) that allow gradients to flow directly through the network and enable training of networks with hundreds of layers. ResNet-based encoders feature in several methods surveyed in Chapter 2, and InceptionResNetV2 serves as the encoder backbone in the FP-U-Net++ architecture [9].

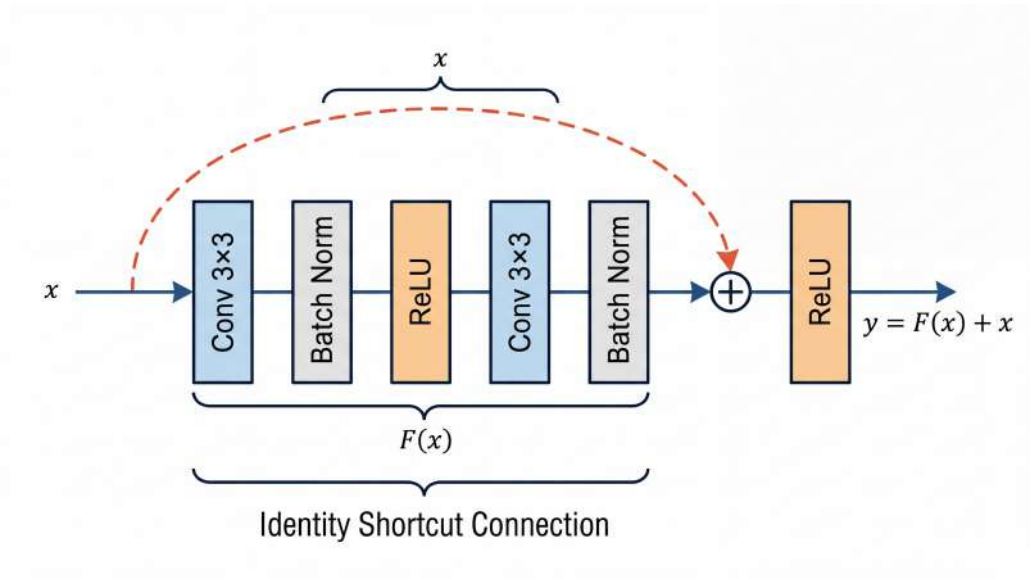


Figure 1.8: Architecture of the ResNet residual block. The identity shortcut connection adds the input $\mathbf{x}$ to the output of the convolutional path $\mathcal{F}(\mathbf{x})$, producing $\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$. This design mitigates the vanishing-gradient problem and enables training of very deep networks [30].

- *U-Net* [5], designed for biomedical image segmentation, has a symmetric encoder–decoder structure with skip connections that concatenate encoder feature maps directly into the corresponding decoder level, enabling precise localisation. U-Net forms the basis of the segmentation decoders in our proposed system (Chapter 3).

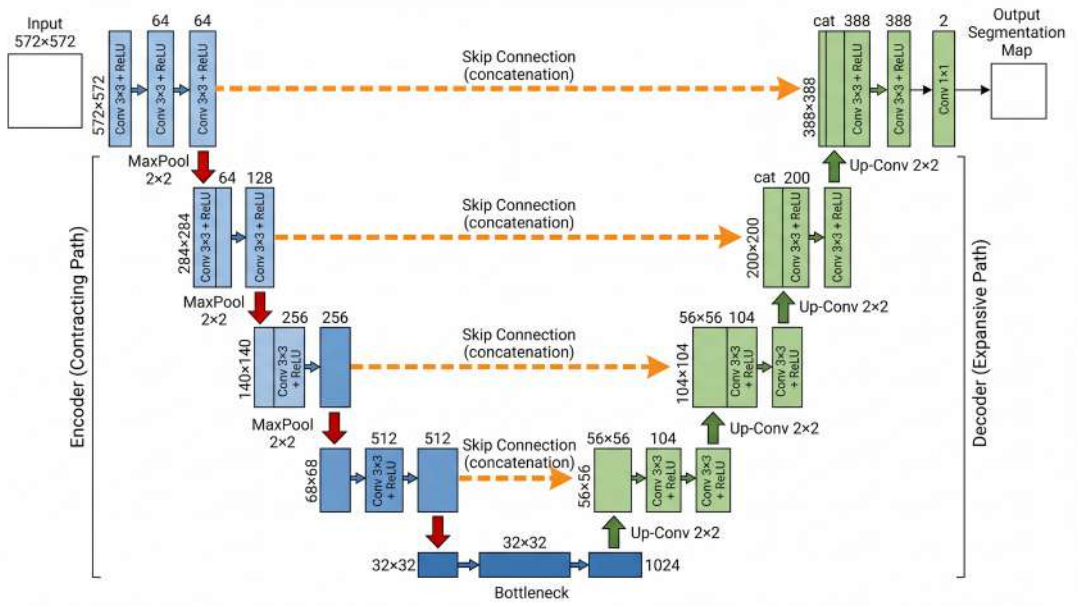


Figure 1.9: The U-Net architecture for biomedical image segmentation. The symmetric encoder (left, contracting path) and decoder (right, expansive path) are linked by skip connections at every resolution level, which concatenate encoder feature maps directly into the corresponding decoder level. This design preserves high-resolution spatial information for precise localisation [5].

- *EfficientNet* [3] is a family of CNN backbones obtained by jointly scaling the network depth, width and input resolution according to a single compound coefficient. The basic block is the mobile inverted bottleneck (MBConv), a depthwise separable convolution wrapped in a residual squeeze-and- excitation structure. EfficientNet variants (B0–B7) attain strong accuracy at comparatively low parameter and FLOP budgets, which makes them attractive transfer-learning backbones in medical imaging (e.g., Gupta et al. [44] use EfficientNetB3 on CCTA, Chapter 2).

1.6.3 Transformer Architectures

Originally proposed for natural language processing, Transformers [31] have been successfully adapted for vision tasks to capture long-range global dependencies that CNNswith their inherently local receptive fieldsmay miss.

Self-Attention Mechanism

Self-attention computes the pairwise relationship between every position in the input sequence [31]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1.3}$$

where Q, K, and V are the Query, Key, and Value matrices respectively, and d_k is the key dimension used for scaling.

Swin Transformer

The Swin Transformer [33] is a hierarchical Vision Transformer that computes self-attention within shifted local windows rather than globally. This approach reduces computational complexity from quadratic to linear in the number of patches, while allowing cross-window connections through the window-shifting mechanism. It is highly effective as a backbone for dense prediction tasks such as vessel segmentation.

Hybrid CNN-Transformer Architectures (MaxViT)

Pure Vision Transformers require very large datasets to learn the inductive biases that are built into convolutional networks (translation equivariance, locality). In medical imaging, where labelled data is scarce, hybrid CNN-Transformer architectures combine the strengths of both paradigms: convolutional layers extract local features in the early stages while attention layers model long-range dependencies in the later stages.

MaxViT [15] is a representative example of this family. Each MaxViT block consists of three sequential operators: an MBConv block (a mobile inverted residual convolution borrowed from EfficientNet) that captures local spatial patterns; a *block attention* layer that performs self-attention within non-overlapping local windows; and a *grid attention* layer that performs self-attention on a sparse grid spanning the entire feature map, yielding global context with linear complexity. This local-then-global design is well suited to coronary artery analysis, where vessel boundaries are local features but the topology of the coronary tree is intrinsically global.

1.6.4 Multi-Task Learning

Multi-task learning (MTL) is a learning paradigm in which a single model is trained simultaneously on several related tasks, sharing representations between them [17, 18]. Compared with training one model per task, MTL has three desirable properties: (i) the shared representation acts as an *implicit regulariser* particularly valuable when each individual task has a small labelled dataset; (ii) features learned for one task can transfer positively to the others (positive task transfer); and (iii) inference becomes faster and cheaper because a single forward pass produces all outputs.

It should be noted that task transfer is not always positive: when the tasks are weakly related or the shared capacity is insufficient, jointly training them can degrade one or more tasks relative to single-task training, a phenomenon known as *negative transfer*. This effect is observed in the multi-source extension of Chapter 3, where auxiliary clinical objectives compete with the imaging tasks on a small single-centre cohort. The reliance on this single-centre cohort is itself a limitation of the approach, which Chapter 3 addresses separately through an external validation of the stenosis branch on the independent public ARCADE dataset.

Hard versus Soft Parameter Sharing

Two main MTL paradigms exist [18]. In *hard parameter sharing*, several tasks share the same encoder and only the final task-specific heads remain separate; this strongly regularises the encoder but assumes the tasks are sufficiently related. In *soft parameter sharing*, each task has its own encoder and a regularisation term encourages the encoders to remain similar; this is more flexible but doubles or triples the parameter count.

Coronary vessel segmentation, stenosis localisation, and QCA regression are strongly anatomically related a stenosis can only exist on a vessel, and QCA quantifies a stenosis which makes them an ideal candidate for hard parameter sharing. This is the design adopted by GCT-Net (Chapter 3).

Multi-Source Learning under Partial Supervision

Multi-task learning concerns *what* a model predicts (several outputs from one shared representation); *multi-source learning* concerns *where the supervision comes from*. In many clinical settings no single label source covers the entire cohort: some patients carry a manually drawn mask, others only machine-generated measurements, and others no annotation at all for a given task. A multi-source model is trained jointly from these heterogeneous and only partially overlapping label sources — for example, manual masks, pseudo-labels distilled from a teacher network (Section 1.6.6), and numeric values parsed from acquisition metadata — rather than from one homogeneous, fully annotated dataset.

The central difficulty is *partial supervision*: each sample carries ground truth for only a subset of the tasks, so the loss must not penalise a head for an output whose target is missing. The standard mechanism, adopted in Chapter 3, attaches a per-sample binary *validity flag* to each task and masks out the loss contributions of absent labels, so that every example contributes only to the tasks for which it carries supervision [39]. Sources of differing reliability can additionally be given different weights (e.g. a manual mask counting fully and a teacher pseudo-label counting at a reduced weight). This is what allows a single GCT-Net to be trained on the heterogeneous label sources of the Oasis cohort; the distinction between the single-source model (Flow 2) and its multi-source extension (Flow 3) is developed in Chapter 3.

Multi-Source Fusion and Feature-wise Linear Modulation (FiLM)

When a model must combine inputs from different modalities — for example, an image, a tabular feature vector, and structured text fields — the modality embeddings have to be fused into a shared representation. A lightweight and widely used fusion mechanism is *Feature-wise Linear Modulation* (FiLM) [34], in which one modality (the *conditioning* signal) produces a per-channel scale γ and shift β that affinely transform the feature map

x of another modality:

$$\mathrm{FiLM}(x) \;=\; \gamma \odot x + \beta, \tag{1.4}$$

where $\odot$ denotes channel-wise multiplication. FiLM conditions the host network on the auxiliary modality while adding only two small projection layers, which makes it far cheaper than full cross-attention fusion and robust to a missing modality (the modulation can degrade gracefully to the identity). This is the mechanism used to inject the tabular ECG embedding into the image bottleneck of the multi-source model in Chapter 3.

Multi-Task Loss Balancing

The simplest way to combine several task losses is the naïve weighted sum

$$\mathcal{L} \;=\; \sum_{t} \lambda_t \, \mathcal{L}_t, \tag{1.5}$$

where $\mathcal{L}$ is the total multi-task training loss, $\mathcal{L}_t$ is the task-specific loss for task t (e.g., vessel segmentation, stenosis localisation, QCA regression), and $\lambda_t \geq 0$ is a scalar weight controlling the contribution of task t to the total. This form requires manually tuning the weights λ_t, which is impractical when tasks have very different scales and noise levels. Kendall et al. [16] proposed *homoscedastic uncertainty weighting*, where each task is assigned a learnable log-variance $\log \sigma_t^2$ and the loss becomes

$$\mathcal{L} \;=\; \sum_{t} \left(\frac{1}{2\sigma_t^2} \, \mathcal{L}_t \;+\; \frac{1}{2} \log \sigma_t^2 \right), \tag{1.6}$$

where $\mathcal{L}$ is again the total multi-task loss, $\mathcal{L}_t$ the task-specific loss for task t, and $\sigma_t > 0$ is a learnable parameter representing the homoscedastic (data-independent) uncertainty associated with task t. Tasks with noisier labels (e.g., QCA values parsed from DICOM tags) automatically acquire a larger σ_t and therefore a smaller effective weight $1/(2\sigma_t^2)$, while the $\log \sigma_t^2$ regularisation prevents the uncertainties from collapsing to zero. This is the loss-balancing strategy used in our system.

1.6.5 Loss Functions

The loss function quantifies the discrepancy between model predictions and ground truth labels, directly governing the optimisation trajectory during training [2]. Different tasks in cardiovascular image analysis require tailored loss formulations.

Binary Cross-Entropy Loss

Binary Cross-Entropy (BCE) is the standard loss for pixel-wise binary classification, including vessel segmentation where each pixel is labelled as vessel or background [5]:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{1.7}$$

where y_i is the ground truth label, $\hat{y}_i$ is the predicted probability, and N is the total number of pixels. BCE treats each pixel independently, which can be problematic when there is severe class imbalance (vessel pixels constitute only 5–10% of an angiographic frame).

Dice Loss

Dice loss directly optimises the Dice coefficient (Section 2.2.2), addressing the class imbalance problem inherent in segmentation tasks [46]:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{1.8}$$

where ϵ is a small smoothing constant to prevent division by zero. Unlike BCE, Dice loss considers the global overlap between prediction and ground truth, making it particularly effective for thin, elongated structures such as coronary arteries.

Focal Loss

Focal loss was introduced by Lin et al. [10] to address extreme foreground-background class imbalance in object detection:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{1.9}$$

where p_t is the predicted probability for the true class, α_t is a class-balancing weight, and $\gamma \geq 0$ is the focusing parameter. When $\gamma = 0$, focal loss reduces to standard cross-entropy. As γ increases (commonly $\gamma = 2$), the loss contribution from well-classified easy examples is down-weighted, forcing the model to focus on hard examples such as small stenotic lesions.

Focal Tversky Loss

The Tversky index generalises the Dice coefficient by introducing separate weights on false positives and false negatives, and the *Focal Tversky* loss of Abraham and Khan [57] adds

a focusing exponent on top:

$$\mathcal{L}_{\mathrm{FT}} = \left(1 - \frac{\sum_i y_i \hat{y}_i + \epsilon}{\sum_i y_i \hat{y}_i + \alpha \sum_i (1 - y_i)\hat{y}_i + \beta \sum_i y_i (1 - \hat{y}_i) + \epsilon}\right)^{1/\gamma}, \tag{1.10}$$

where α and β weight false positives and false negatives respectively (with $\alpha + \beta = 1$) and γ is the focusing exponent. Setting $\alpha > \beta$ penalises false negatives more heavily, which is desirable for small-lesion segmentation such as stenosis localisation; the focusing exponent concentrates the gradient on hard, low-overlap cases. This is the primary stenosis-segmentation loss used in Chapter 3.

Compound Loss Functions

Many state-of-the-art methods combine multiple losses to exploit their complementary strengths [9]:

$$\mathcal{L}_{\mathrm{total}} = \lambda_1 \mathcal{L}_{\mathrm{BCE}} + \lambda_2 \mathcal{L}_{\mathrm{Dice}} + \lambda_3 \mathcal{L}_{\mathrm{reg}} \tag{1.11}$$

where $\mathcal{L}_{\mathrm{reg}}$ denotes a regularisation term (e.g., L2 weight decay) and $\lambda_1, \lambda_2, \lambda_3$ are weighting hyperparameters.

Boundary (Sobel) Loss

Overlap-based losses such as Dice and Focal Tversky reward area agreement but are relatively insensitive to the precise location of object boundaries a problem when the structures of interest are thin, elongated vessels or small stenotic lesions whose clinical interpretation depends critically on edge accuracy. A simple way to inject an explicit boundary supervision is to compute the spatial gradient of both the prediction and the ground truth with a Sobel filter and penalise the squared difference of the resulting gradient magnitudes:

$$\mathcal{L}_{\mathrm{Bnd}} = \frac{1}{N} \sum_{i=1}^{N} \big(\|\nabla \hat{y}_i\|_2 - \|\nabla y_i\|_2 \big)^2, \tag{1.12}$$

where ∇ is the discrete gradient operator implemented as two depthwise convolutions with the horizontal and vertical 33 Sobel kernels. In our system this term is added to the stenosis-segmentation objective to sharpen lesion contours.

Topology-Preserving Centerline-Dice Loss (clDice)

Volumetric Dice rewards pixelwise area overlap but is insensitive to the *connectivity* of the predicted structure: a thin vessel can be correctly painted in terms of area while being broken into disconnected fragments. For tubular anatomies such as the coronary tree, where the clinical relevance of a segmentation depends on continuous tracing along the vessel centreline, this is a meaningful failure mode. The *centerline-Dice* (clDice) loss

introduced by Shit et al. [55] couples the volumetric prediction with its skeleton in order to enforce topological correctness. Let $S(\hat{y})$ and $S(y)$ denote the soft skeletons of the prediction and the ground truth, obtained by an iterative differentiable erode–dilate–residual operator. Two scores are defined:

$$\mathrm{T}_{\mathrm{prec}} = \frac{|S(\hat{y}) \cap y|}{|S(\hat{y})|}, \qquad \mathrm{T}_{\mathrm{sens}} = \frac{|S(y) \cap \hat{y}|}{|S(y)|}, \tag{1.13}$$

the first measuring the fraction of the predicted skeleton lying inside the ground-truth volume and the second measuring the fraction of the ground-truth skeleton lying inside the predicted volume. The clDice loss is the harmonic-mean complement

$$\mathcal{L}_{\mathrm{clDice}} \;=\; 1 - \frac{2 \cdot \mathrm{T}_{\mathrm{prec}} \cdot \mathrm{T}_{\mathrm{sens}}}{\mathrm{T}_{\mathrm{prec}} + \mathrm{T}_{\mathrm{sens}}}. \tag{1.14}$$

Because the soft skeletonisation is differentiable, the loss can be back-propagated through. In our system clDice is applied to the vessel head only, and only on samples carrying *manual* masks (not on teacher pseudo-labels), in order to avoid corrupting the topology objective with label noise.

Conformal Prediction for Regression Outputs

Among the metrics used to evaluate a continuous regression output, the *Mean Absolute Error* (MAE) is the simplest. Given N test samples with ground-truth values $\{y_i\}_{i=1}^{N}$ and predictions $\{\hat{y}_i\}_{i=1}^{N}$, it is defined as

$$\mathrm{MAE} \;=\; \frac{1}{N} \sum_{i=1}^{N} \big|\, \hat{y}_i - y_i \,\big|, \tag{1.15}$$

and is expressed in the same physical unit as the target. MAE summarises central error but does not communicate prediction uncertainty: a model with MAE = 11 percentage points on $\%DS$ may make ±5 pp errors most of the time and ±30 pp errors on a minority of cases. Where a learned regressor is consumed in a clinical decision support context, an *interval* estimate that covers the true value with a controlled probability is preferable. *Split-conformal prediction* [56] is a distribution-free wrapper that turns any pretrained point predictor into an interval predictor with a marginal coverage guarantee. After training, a held-out calibration set $\{(x_i, y_i)\}_{i=1}^{n}$ is used to compute non-conformity scores $s_i = |\hat{f}(x_i) - y_i|/\sigma(x_i)$, where $\sigma(x_i)$ is an optional input-dependent normaliser. Given a target miscoverage $\alpha \in (0, 1)$, the conformal quantile is

$$\hat{q} \;=\; \mathrm{Quantile}\!\left(\{s_i\}_{i=1}^{n};\; \frac{\lceil (1-\alpha)(n+1) \rceil}{n} \right), \tag{1.16}$$

and the prediction interval at a new point x is $[\hat{f}(x) - \hat{q}\,\sigma(x),\ \hat{f}(x) + \hat{q}\,\sigma(x)]$. Under the exchangeability assumption that calibration and test points are drawn from the same distribution, this interval covers the true y with probability at least $1-\alpha$, regardless of the underlying model and without distributional assumptions on the residuals. The choice of $\sigma(x)$ governs how the interval width *adapts* to the difficulty of each input: a larger $\sigma(x)$ on harder inputs produces wider intervals on those inputs while preserving the marginal coverage guarantee.

Smooth L1 (Huber) Loss

For continuous regression targets such as the stenosis percentage and lesion length predicted by the QCA head pure L2 loss is sensitive to outliers, while pure L1 loss is non-smooth at zero and slows convergence. The *Smooth L1 (Huber) loss* [38] combines the two: it is quadratic for small errors (stable gradient near the optimum) and linear for large errors (robust to noisy labels). With threshold β, it is defined as

$$\mathcal{L}_{\text{Huber}}(\hat{y}, y) = \begin{cases} \dfrac{1}{2\beta}(\hat{y} - y)^2 & \text{if } |\hat{y} - y| < \beta \\ |\hat{y} - y| - \dfrac{\beta}{2} & \text{otherwise.} \end{cases} \tag{1.17}$$

where $\hat{y}$ and y are the predicted and ground-truth values and the threshold $\beta > 0$ is the error magnitude at which the loss switches from its quadratic (L2-like) regime, used for small residuals, to its linear (L1-like) regime, used for large residuals. This is the loss adopted for QCA regression in our work, where the targets are parsed from DICOM private tags and may contain occasional noisy values.

1.6.6 Transfer Learning

Transfer learning is a strategy where a model pre-trained on a large-scale source task is adapted to a different target task [59]. In deep learning, this typically involves, for example, taking a network pre-trained on ImageNet [35] (over 14 million natural images across 1,000 categories) and fine-tuning it for a specialised domain such as medical imaging, where expert-annotated data is scarce. For example, a ResNet trained to classify the 1,000 categories of natural images in ImageNet can be re-used, with its final classification layer replaced and its earlier convolutional layers fine-tuned, as the encoder of a coronary-vessel segmentation network — transferring the low- and mid-level visual features (edges, textures, contour shapes) that generalise across domains.

In practice, transfer learning is implemented through two main approaches:

- *Feature extraction:* the pre-trained convolutional base is frozen, and only a newly added task-specific head is trained. This is suitable when the target dataset is very small.

- *Fine-tuning:* the entire network is trained end-to-end on the target dataset with a reduced learning rate, allowing the pre-trained features to gradually adapt to the new domain.

Transfer learning is particularly critical in medical imaging, where expert-annotated datasets are often limited in size — the dataset used in this work, for example, comprises only 200 patients. By leveraging pre-trained weights from ImageNet [35], models such as ResNet and InceptionResNetV2 can achieve strong performance on coronary artery analysis tasks without requiring millions of medical training images [9, 44]. The same fine-tuning principle is exploited again at the level of *datasets* rather than tasks: in Chapter 3, the model trained on the single-centre Oasis cohort is used as an initialisation and fine-tuned on a second, independently acquired public dataset (ARCADE) to measure how well the learned representation transfers across acquisition domains.

Knowledge Distillation and Pseudo-Labelling

When ground-truth labels are unavailable for part of a dataset, a closely related strategy is *knowledge distillation* [36], in which a larger or earlier-trained "teacher" network produces soft predictions that are used as supervisory targets for a "student" network. In segmentation problems, this typically takes the form of *pseudo-labelling*: the teacher's sigmoidal output map is treated as a probabilistic mask, and the student is trained to reproduce it on unlabelled samples, often with a lower loss weight to reflect the uncertainty of the pseudo-target. In this work, an existing U-Net trained on a fully supervised vessel-segmentation dataset acts as the teacher and provides soft vessel pseudo-labels for clinical samples that lack manual vessel annotation, allowing the proposed multi-task model to be trained on a much larger effective training set than manual annotation alone would permit (Chapter 3).

Self-Supervised Learning and Masked Image Modeling

Self-supervised learning (SSL) pretrains a network on a pretext task derived from the unlabelled data itself, learning useful representations without manual annotation. A prominent SSL approach for vision is *masked image modeling* (MIM), exemplified by the Masked Autoencoder (MAE) [58]: a large fraction of image patches is masked out and the network is trained to reconstruct the missing content from the visible patches, forcing it to learn the structure of the image. MIM provides an alternative to ImageNet transfer learning for initialising an encoder when domain-specific unlabelled data is abundant; a vessel-aware variant is discussed as future work for encoder pretraining in Chapter 3.

1.6.7 Data Augmentation

Data augmentation artificially increases the size and diversity of the training set by applying random transformations to existing samples and is essential in medical imaging where labelled data is limited. Common augmentation strategies include:

- *Geometric transformations* such as random rotation, horizontal/vertical flipping, and scaling, which simulate variations in patient anatomy and imaging angle.
- *Elastic and grid distortions:* non-rigid spatial deformations that locally warp the image with a smooth random displacement field. They simulate anatomical variability between patients and the small positional differences observed across angiographic projections, a relevant prior for coronary trees whose morphology differs from one patient to another.
- *Intensity transformations* such as random brightness, contrast adjustment and Gaussian noise, simulating variations in imaging equipment and contrast-agent concentration.
- *Spatial transformations* such as random cropping and padding, which encourage the model to learn from partial views and reduce overfitting to specific spatial positions.

In our implementation, augmentations are applied through the `Albumentations` library [20], which guarantees that the same geometric transform is applied to the input image and to all of its associated masks (vessel mask, stenosis mask), preserving pixel-wise alignment.

Data augmentation acts as an implicit regulariser, reducing overfitting and improving the model's ability to generalise to unseen clinical cases.

Test-Time Augmentation (TTA)

Augmentation is not limited to the training phase. *Test-time augmentation* (TTA) consists of running inference on several transformed versions of the input image (e.g., the original image and its three flipped variants), inverting the geometric transformation on each output, and averaging the results. The averaged prediction is more robust than any single forward pass because it suppresses the variance introduced by spatial position and orientation. This strategy is used in the inference pipeline of our proposed system (Chapter 3).

1.6.8 Mixed-Precision Training

Modern GPUs (NVIDIA Pascal and later) execute half-precision floating-point (FP16) operations significantly faster than full-precision (FP32) operations and consume less

memory. *Mixed-precision training* [37] keeps a master copy of the weights in FP32 but performs forward and backward passes in FP16, with a *loss scaler* that multiplies the loss by a large constant before back-propagation to prevent gradient underflow, then unscales the gradients before the optimiser update. In practice this halves the GPU memory footprint and speeds up training by roughly 1.5 to 2 on modern hardware, which is essential when training a ∼30M-parameter model on a 512512 multi-task pipeline within a single Tesla P100 GPU.

1.7 Conclusion

This chapter has provided the conceptual background for the work that follows. It first reviewed heart and coronary anatomy and the clinical aspects of coronary artery disease, including stenosis grading, treatment strategies, and medical reporting. The three diagnostic modalities employed in this work — ECG, CCTA, and ICA — were then introduced alongside the DICOM standard and the digital image processing techniques used in the preprocessing pipeline.

The chapter then established the deep learning foundations relevant to the proposed system. Convolutional and Transformer architectures were presented in turn, followed by the hybrid CNN-Transformer family illustrated through MaxViT. The multi-task learning paradigm was introduced next, covering hard and soft parameter sharing, FiLM-based multi-task fusion, and homoscedastic uncertainty weighting for loss balancing. The complementary notion of multi-source learning under partial supervision — training a single model from heterogeneous, partially overlapping label sources via per-sample validity flags — was introduced as it underpins the multi-source extension of the proposed system.

The task-specific loss functions used in this work were then detailed, including Dice, Focal Tversky, the Sobel boundary term, topology-preserving centerline-Dice, and Smooth L1, together with the principles of distribution-free conformal prediction for regression interval estimation. Finally, the supporting training machinery was described: transfer learning, knowledge distillation, data augmentation, test-time augmentation, and mixed-precision training. These elements form the basis of the proposed multi-task system, presented in Chapter 3.

Chapter 2

Deep Learning for Cardiovascular Disease Detection: State of the Art

2.1 Introduction

Regarding the abundance of work in the field, this chapter presents a review of recent deep learning techniques applied to cardiovascular disease detection. The reviewed methods are organised into three categories: vessel segmentation, lesion detection, and stenosis quantification. For each category, this section analyses the deep learning architectures employed, the datasets used, the results reported, and the limitations encountered. The survey focuses solely on deep learning-based approaches, and sets aside traditional and machine learning techniques.

This three-part taxonomy reflects the natural hierarchical structure of the coronary analysis pipeline. Vessel segmentation is the foundational step: the coronary tree must first be delineated before any pathology can be located. Stenosis detection builds on this foundation by identifying and localising abnormal narrowings along the segmented vessel. Stenosis quantification is the terminal step: once a lesion is localised, its severity can be expressed in objective clinical terms — percent diameter stenosis (%DS) and lesion length — to guide the decision on revascularisation. Each stage is therefore a strict prerequisite for the next, and a failure at one stage propagates directly to those that follow. Organising the literature along these three axes allows every reviewed method to be positioned precisely within the analysis chain. This structure also directly anticipates the three task-specific heads of GCT-Net — vessel segmentation, stenosis localisation, and QCA regression — presented in Chapter 3.

2.2 Datasets and Evaluation Metrics

2.2.1 Datasets

Deep learning models for cardiovascular disease detection rely on a variety of imaging datasets, which can be categorised according to their imaging modality, annotation type, and primary task. Table 2.1 summarises the most commonly used publicly available datasets in this field. The datasets cover three main modalities: X-ray invasive coronary angiography, represented by two databases, ARCADE and CADICA, coronary CT angiography (ASOCA), and ECG signals (Kaggle ECG). The annotation granularity ranges from binary classification labels (normal/abnormal) in ECG datasets, to pixel-level masks for vessel segmentation (ARCADE, ASOCA), to semi-quantitative stenosis percentage categories (CADICA). In addition, two single-centre ICA datasets by Zhao et al. are widely used as supervised benchmarks for combined segmentation and quantification tasks.

Table 2.1: Publicly available datasets for CVD detection using deep learning.

Dataset	**Size**	**Modality**	**Annotation**
ARCADE (2023) [47, 49, 50]	1,500 images	X-ray Angiography	COCO format, stenosis masks
CADICA (2025) [7]	350 videos	ICA Videos	Stenosis percentage categories
ASOCA (2024) [45]	Variable	CT Angiography	Coronary artery segmentation
Kaggle ECG (2025) [42]	4,997 records	ECG signals	Binary (Normal/Abnormal)
Zhao et al. (2020) ICA [46]	294 images (73 patients)	non-invasive Coronary Angiography	Manual vessel contours, stenosis levels
Zhao et al. (2021) ICA [9]	314 images (99 patients)	non-invasive Coronary Angiography	Manual vessel contours, stenosis levels

2.2.2 Evaluation Metrics

Rigorous evaluation is critical in medical AI, where model performance has direct patient-safety implications. The metrics used in the studies reviewed in this chapter fall into three groups: *classification metrics* (used for ECG-based detection and stenosis severity grading), *segmentation metrics* (used for vessel delineation and lesion boundary extraction), and *task-specific measures* employed in object detection and stenosis quantification.

Classification Metrics

For binary or multi-class classification, predictions are categorised against the ground truth as True Positives (TP), True Negatives (TN), False Positives (FP), or False Negatives (FN). The standard derived metrics are:

- *Accuracy:* the proportion of correct predictions over the total number of samples,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.1}$$

- *Sensitivity (recall):* the ability to correctly identify diseased cases. Missing a positive case (a False Negative) carries severe clinical consequences, which is why recall is often the primary metric in cardiovascular detection tasks,

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.2}$$

- *Specificity:* the ability to correctly identify healthy cases,

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{2.3}$$

- *Precision:* the proportion of positive predictions that are correct,

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.4}$$

- *F1-score:* the harmonic mean of precision and recall, providing a single balanced measure that is useful when classes are imbalanced,

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.5}$$

- *AUC-ROC:* the Area Under the Receiver Operating Characteristic curve, assessing discriminative performance across all classification thresholds. A perfect classifier yields AUC= 1, a random classifier AUC= 0.5.

Segmentation Metrics

Segmentation performance is evaluated using overlap-based metrics that measure the spatial agreement between the predicted mask P and the ground-truth mask G:

- *Intersection over Union (IoU)*, also known as the Jaccard index,

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}. \tag{2.6}$$

- *Dice coefficient*, the most common metric for medical image segmentation, particularly suited to thin elongated structures such as coronary arteries,

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}. \tag{2.7}$$

Task-Specific Metrics

In addition to the standard classification and segmentation metrics, several task-specific measures are used in the studies reviewed in this chapter:

- *Mean Average Precision (mAP):* the standard metric for object detection, computed as the mean of Average Precision (AP) values across all classes [41]. *mAP@0.50* uses a single IoU threshold of 0.50, while *mAP@0.50:0.95* averages over thresholds from 0.50 to 0.95 in steps of 0.05, giving a stricter localisation evaluation.
- *True Positive Rate (TPR):* equivalent to recall and used in the stenosis-detection context to measure the proportion of actual stenotic segments correctly identified.
- *Positive Predictive Value (PPV):* equivalent to precision, measuring the proportion of detected stenoses that are true positives.
- *Root Mean Square Error (RMSE):* used for quantification tasks to measure the deviation between predicted and reference stenosis percentages,

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{2.8}$$

  where N is the number of evaluated segments, y_i the ground-truth stenosis percentage and $\hat{y}_i$ the corresponding model prediction.
- *Hausdorff Distance (HD):* the maximum distance between the boundary of the predicted segmentation P and the ground-truth boundary G [9],

$$H(P, G) = \max\left(\sup_{p\in\partial P}\inf_{g\in\partial G}\|p - g\|,\ \sup_{g\in\partial G}\inf_{p\in\partial P}\|g - p\|\right), \tag{2.9}$$

  where ∂P and ∂G denote the boundary sets of P and G. A lower HD indicates tighter spatial agreement at the vessel contour.
- *Brier Score:* a calibration metric measuring the mean squared difference between predicted class probabilities and actual binary outcomes [44],

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2, \tag{2.10}$$

where $\hat{p}_i$ is the predicted probability for the i-th sample and y_i the binary outcome. Lower values indicate better-calibrated models.

2.3 Literature Review

2.3.1 Vessel Segmentation

Liu et al. [47] proposed YOLO-Angio, a three-stage approach for coronary artery tree segmentation that ranked 3rd in the ARCADE Challenge MICCAI 2023. The methodology combines seven image preprocessing techniques with an ensemble of YOLOv8 models and a graph-based anatomical reconstruction post-processing step. On the ARCADE Challenge 2023 dataset, the method achieved F1-score of 0.429, representing a 58% improvement over a single model without post-processing. Limitations include inability to exploit temporal frame information and suboptimal real-time performance.

Liu et al. [8] developed AI-QCA, a fully automatic QCA workflow using BasNet for segmentation and PSPNet for root node location. On 13,222 angiograms from 3,275 patients, the model achieved segmentation F1-score of 0.879 and QCA RMSE of 0.064. Despite its large training set, the system operates as a three-stage sequential pipeline — segmentation, tree construction, then diameter fitting — meaning errors in the segmentation stage propagate directly into the QCA output without correction.

Liu et al. [48] developed the Progressive Perception Learning (PPL) framework incorporating context, interference, and boundary perception modules. Evaluated on 1,086 subjects, the framework achieved Dice score exceeding 0.95, providing the segmentation foundation for AI-QCA. However, the framework addresses vessel segmentation only and does not extend to stenosis detection or quantification, limiting its standalone clinical utility.

Alirr et al. [45] proposed a ResUNet framework combining Frangi vesselness filtering with heart ROI extraction [40]. On the ASOCA and MM-WHS datasets, the method achieved Dice Similarity Coefficient of 0.867, Jaccard Index of 0.765, recall of 0.881, and precision of 0.892, outperforming 3D-UNet (DSC 0.837) and UNETR (DSC 0.827). The method is evaluated exclusively on CT angiography; its applicability to X-ray ICA, which presents lower contrast and higher noise levels, remains untested.

Zhao et al. [46] proposed MIMS U-Net with a two-stage recurrent training strategy on 294 ICAs from 73 patients, achieving a Dice score of 0.833, sensitivity of 0.828, and specificity of 0.998. While the segmentation is learned end-to-end, the subsequent stenosis detection relies on a separate post-hoc geometric algorithm that extracts centrelines and calculates vessel diameters, rather than a jointly trained component.

Zhao et al. [9] proposed FP-U-Net++, integrating feature pyramid supervision into U-Net++ with an InceptionResNetV2 encoder. On 314 ICAs from 99 patients, the model

achieved a Dice score of 0.890, sensitivity of 0.860, specificity of 0.996, and Hausdorff distance of 5.83 mm. Feature pyramid supervision improves vessel delineation accuracy, but stenosis quantification is still derived by a separate streamline algorithm rather than being learned end-to-end alongside the segmentation objective.

Table 2.2 compares these vessel segmentation methods across key metrics.

Table 2.2: Comparative Analysis of Deep Learning Methods for Vessel Segmentation. SN: Sensitivity; SP: Specificity; XRA: X-ray angiography; CTA: CT angiography; ICA: non-invasive coronary angiography.

Study	**Architecture**	**Dataset**	**Modality**	**Dice/F1**	**SN**	**SP**
Liu et al. [47]	YOLOv8 Ensemble	ARCADE (1500)	XRA	0.429	–	–
Liu et al. [8]	BasNet + PSP-Net	Fuwai (13222)	XRA	0.879	–	–
Liu et al. [48]	PPL Framework	1086 subjects	XRA	>0.95	–	–
Alirr et al. [45]	ResUNet	ASOCA + MM-WHS	CTA	0.867	0.881	–
Zhao et al. (2020) [46]	MIMS U-Net	294 ICAs (73 pts)	ICA	0.833	0.828	0.998
Zhao et al. (2021) [9]	FP-U-Net++	314 ICAs (99 pts)	ICA	0.890	0.860	0.996

2.3.2 Lesion Detection

Abrar et al. [42] proposed an automated heart disease detection system using the Swin Transformer architecture [33] for ECG-based classification. The study converts raw one-dimensional ECG signals into four-dimensional representations suitable for transformer-based models, leveraging shifted window self-attention mechanisms to capture both local and global dependencies that traditional CNN and RNN architectures struggle to model. The architecture employs a ResNet152 backbone trained for 200 epochs with batch size 16 and learning rate 0.005 on the Kaggle ECG Heartbeat Dataset containing 4,997 records split into 70% training, 10% validation, and 20% testing. The reported metrics were 99.8% accuracy, 99.72% precision, 99.91% recall, 99.81% F1-score and 99.99% AUC, outperforming the conventional machine learning techniques. Reported limitations include high computational requirements, limited explainability, and the need for validation across diverse ECG patterns and multi-source datasets.

Pokhrel et al. [49] developed a ConvNeXt-V2 backbone integrated with Mask R-CNN [6] for instance segmentation of stenotic lesions in X-ray coronary angiography. The methodology employs Global Response Normalization (GRN) layers to reduce redundant

activations while amplifying feature diversity, with a Regional Proposal Network (RPN) identifying Regions of Interest. Training was conducted for 36 epochs using AdamW optimiser with initial learning rate of 110^{-4}, batch size 8. On the ARCADE dataset, the model achieved an F1-score of 0.535 for stenosis detection, outperforming YOLOv8 (F1 = 0.232), Rtmdet-ins-large (0.265), ResNet-50 Mask R-CNN (0.381), and ResNet-101 Mask R-CNN (0.417).

Ansari [50] evaluated state-of-the-art object detection models including Grounding DINO, YOLOv3, and DINO-DETR on the ARCADE dataset using the MMDetection framework. On 1,500 images with COCO format annotations, DINO-DETR achieved mAP@0.50:0.95 of 0.086 with mAP@0.50 of 0.228, Grounding DINO reached 0.080 with the highest mAP@0.50 of 0.259, while YOLOv3 attained mAP@0.50:0.95 of 0.068 with mAP@0.50 of 0.254. All models struggled significantly with small-scale stenotic lesions (mAP_small < 0.20).

Li et al. [7] proposed LT-YOLO, a long-term temporal enhanced YOLO method for automatic stenosis detection in non-invasive coronary angiography . The architecture integrates long-term temporal information using state-space models (Mamba) combined with YOLOv8 [51]. Evaluated on the CADICA dataset containing 350 ICA videos from 42 patients using 5-fold cross-validation, LT-YOLO achieved mAP@0.50 of 0.764, representing a 2.9% improvement over YOLOv8 baseline (mAP@0.50 = 0.735), with particularly strong gains for moderate stenoses (AP_p50-70 of 0.804 versus 0.759).

Wang et al. [43] developed a two-stage, multiview deep learning model for automated assessment of coronary stenosis in heavily calcified plaques (Agatston score $\geq$300) on CCTA. The first stage employs a Faster R-CNN [38]-based calcification detection network achieving 95% sensitivity at 0.25 false positives per patient, while the second stage uses a multiview ResNet to predict stenosis severity. On 10,101 CCTA examinations externally validated on 442 CCTAs, the model achieved segment-level AUC of 0.89 and vessel-level AUC of 0.90.

Gupta et al. [44] developed end-to-end deep learning models to detect moderate ($\geq$50%) and severe ($\geq$70%) stenosis using curved multiplanar reformations from CCTA scans, employing transfer learning with EfficientNet variants [59]. On approximately 900 patients, EfficientNetB3 achieved AUC-ROC of 0.95 for LAD moderate stenosis.

Zhao et al. [46] developed an automatic arterial stenosis detection algorithm following vessel segmentation in non-invasive coronary angiograms. After segmenting arteries using MIMS U-Net, the method achieved a TPR of 0.667, PPV of 0.704, and RMSE of 0.174 (see Equation 2.8).

Zhao et al. [9] further improved stenosis detection in terms of TPR using FP-U-Net++ and achieved a TPR of 0.684, PPV of 0.700, ARMSE of 0.154, and RRMSE of 0.220, outperforming all competing methods including DeepLabV3 (0.318), AngioNet (0.633), U-Net (0.663), and U-Net++ (0.651).

As summarised in Table 2.3, transformer-based and CNN-hybrid architectures dominate lesion detection performance across modalities.

Table 2.3: Comparative Analysis of Deep Learning Methods for Lesion Detection.

Study	**Architecture**	**Dataset**	**Modality**	**Best Metric**	**Task**
Abrar et al. [42]	Swin Transformer	Kaggle ECG	ECG	Acc = 0.998	ECG Classification
Pokhrel et al. [49]	ConvNeXt-V2 + Mask R-CNN	ARCADE	XRA	F1 = 0.535	Stenosis Segmentation
Ansari [50]	Grounding DINO	ARCADE	XRA	mAP@0.50 = 0.259	Stenosis Detection
Li et al. [7]	LT-YOLO (Mamba)	CADICA	ICA Video	mAP@0.50 = 0.764	Video Stenosis
Wang et al. [43]	Faster R-CNN + ResNet	CCTA	CT	AUC = 0.89	Calcified Plaque
Gupta et al. [44]	EfficientNetB3	CCTA (CMR)	CT	AUC = 0.95	Stenosis Detection
Zhao et al. (2020) [46]	MIMS U-Net + Algorithm	294 ICAs	ICA	TPR = 0.667	Stenosis Detection
Zhao et al. (2021) [9]	FP-U-Net++ + Algorithm	314 ICAs	ICA	TPR = 0.684	Stenosis Detection

2.3.3 Lesion and Stenosis Quantification

Abrar et al. [42] provide quantification through binary classification of ECG signals. The quantification is performed via a SoftMax output layer that yields probability scores for Normal versus Abnormal classes, achieving AUC of 0.99.

Wang et al. [43] perform stenosis quantification as a regression task: the multi-view ResNet outputs a continuous stenosis percentage from curved planar reconstruction patches, which is then thresholded at $\geq$50% to classify segments as obstructive. The method achieves segment-level AUC of 0.89 and 96% CAD-RADS categorical agreement.

Liu et al. [8] automatically extract QCA parameters using a centreline-based approach: after segmentation, the vessel centreline is extracted via Zhang's thinning algorithm; diameters are computed using Gaussian-smoothed cross-sectional width measurement; and the stenosis diameter is localised by comparing the profile to an interpolated reference diameter. The system achieves RMSE of 0.064 for diameter stenosis versus manual QCA (Pearson $r = 0.765$).

Zhao et al. [46] compute the stenotic level for each arterial segment as $s = (1 - d_{\text{min}}/d_{\text{ref}})100\%$, where d_{min} is the minimum lumen diameter and d_{ref} is the reference (maximum) diameter along the centreline extracted via morphological erosion and the Euclidean distance transform. The method achieved TPR of 0.667, PPV of 0.704, and RMSE of 0.174.

Zhao et al. [9] further improved quantification using FP-U-Net++ contours and backward difference methods to detect local diameter extrema, achieving TPR of 0.684, PPV of 0.700, ARMSE of 0.154, and RRMSE of 0.220. Particularly noteworthy is the improvement on severe stenosis (TPR 0.556 vs. 0.333 for U-Net and U-Net++), which is the most clinically significant grade.

A quantitative comparison of these methods is provided in Table 2.4.

Table 2.4: Comparative Analysis of Deep Learning Methods for Lesion/Stenosis Quantification.

Study	**Architecture**	**Quantification Method**	**Best Metric**	**Output**
Abrar et al. [42]	Swin Transformer	Softmax classification	Acc = 0.998, AUC = 0.9999	Normal/Abnormal probability
Wang et al. [43]	Faster R-CNN + ResNet	Stenosis regression	AUC = 0.89 (segment)	Stenosis % + CAD-RADS
Liu et al. [8]	BasNet + PSPNet	Centreline-based QCA	RMSE = 0.064 (DS)	LL, MLD, RVD, DS
Zhao et al. (2020) [46]	MIMS U-Net	Diameter-based centreline	TPR = 0.667, RMSE = 0.174	Stenosis % (4 grades)
Zhao et al. (2021) [9]	FP-U-Net++	Backward-difference diameter	TPR = 0.684, ARMSE = 0.154	Stenosis % (4 grades)

2.3.4 Multi-Task Approaches in Coronary and Vessel Analysis

The studies discussed so far address vessel segmentation, lesion detection and stenosis quantification as separate problems, typically solved by sequential pipelines. A complementary line of research formulates two or more of these sub-tasks within a single network and is directly relevant to the present work.

Tetteh et al. [23] introduced DeepVesselNet, a multi-task framework that performs vessel segmentation, centreline prediction and bifurcation detection on 3D angiographic volumes (MRA, CTA) within a shared backbone. The work explicitly motivates multi-task design on the grounds that "moving from raw angiography images to vessel segmentation alone might not provide enough information for clinical use" and constitutes an early instance of hard parameter sharing for vascular analysis.

Lin et al. [22] reported a multi-centre study on coronary CCTA in which the segmentation network was trained in a multi-task configuration over two related anatomical objectives ((i) the vessel wall and (ii) the lumen and plaque components) to enforce structural consistency between predictions. The downstream stenosis percentage was then derived from the segmented lumen geometry rather than predicted directly by a regression head.

Yao et al. [21] proposed MGFA-Net, a dual-encoder U-shaped architecture for joint coronary artery segmentation and stenosis assessment on CCTA, integrating myocardial-region prior knowledge as a guidance signal and quantifying prediction uncertainty via

Monte Carlo dropout. Stenosis grading is obtained by applying a centreline-based morphological algorithm to the predicted segmentation, rather than by a learned regression head; the closest similarity to the present work is the joint segmentation–assessment formulation, while the modality (CCTA), the quantification mechanism (post-hoc geometric versus learned regression) and the targeted clinical metrics differ.

In a related direction, Park et al. [52] described an artificial-intelligence QCA system trained on 7,658 angiographic images that uses three deep learning models for lumen-boundary delineation, with geometric measurements (%DS, MLD, RVD, LL) computed from the resulting contours. The system reaches a per-lesion sensitivity of 89% on a European validation cohort but, like AI-QCA [8], it is a multi-stage pipeline rather than a single end-to-end multi-task network.

Huang et al. [28] similarly combine a SAM-based and VM-UNet segmentation backbone with a downstream centreline-based stenosis detection algorithm. Du et al. [53] likewise report a three-stage end-to-end pipeline (key-frame extraction, vessel segmentation, stenosis measurement) for ICA sequences, in which the three stages are trained independently rather than as a single multi-task objective.

For the choice of architecture, Gerbasi et al. [27] introduced the first Multi-Axis Vision Transformer (MA-ViT) for coronary imaging, applying it to CAD-RADS scoring on CCTA with reported AUC values of 0.87–0.93. This work establishes Multi-Axis attention as a competitive backbone family in coronary analysis. The MaxViT backbone [15] adopted in this thesis is from the same family.

For the loss-balancing mechanism, Liu et al. [25] proposed AHU-MultiNet, a three-task medical image segmentation network in which homoscedastic uncertainty weighting [16] balances a primary segmentation objective with two auxiliary tasks (signed distance regression and contour detection). Evaluated on the 2018 Atrial Segmentation Challenge and the 2017 Liver Tumour Segmentation Challenge, the multi-task formulation with adaptive uncertainty weighting consistently outperformed both the single-task baseline and uniform-weight multi-task variants, confirming that jointly learned uncertainty weights improve segmentation over separately trained models.

Bragman et al. [26] previously applied the same uncertainty weighting to a joint segmentation–regression problem in MR-only radiotherapy planning, simultaneously segmenting organs-at-risk and regressing a synthetic CT from MRI. The multi-task model achieved a synCT MAE of $\sim$73.9 HU, compared to 92.9 HU for a single-task 3D U-Net and 89.1 HU for HighRes3DNet, while also reaching state-of-the-art organ segmentation — a direct demonstration that MTL with uncertainty weighting outperforms separate single-task pipelines on a mixed segmentation–regression objective. Both works demonstrate that Kendall's formulation transfers well from its original scene-understanding setting [16] to medical imaging tasks of mixed segmentation and regression nature, which is the regime in which GCT-Net operates.

Finally, on the supervision side, the present work draws on partially supervised medical-image segmentation, in which images carry only a subset of the labels of interest. Reiß et al. [39] showed that a single segmentation network can be trained from images with heterogeneous annotation strengths (full masks, bounding boxes, image-level tags, or no labels at all) by using validity-aware deep-supervision losses, an approach conceptually similar to the per-sample validity-flag mechanism employed in the present work.

Table 2.5 summarises the principal multi-task and joint-quantification approaches reviewed in this subsection.

Table 2.5: Multi-task and joint quantification approaches in coronary and vascular analysis.

Study	Modality	Architecture	Tasks	Quantification
Tetteh et al. [23]	3D MRA / CTA	3D CNN, shared backbone	Vessel seg. + centreline + bifurcation	None (vessel-tree only)
Lin et al. [22]	CCTA	ConvLSTM, multitask seg.	Vessel wall seg. + lumen/plaque seg.	Geometric, post-hoc
Yao et al. [21]	CCTA	Dual-encoder U-Net	Coronary seg. + stenosis assessment	Centreline-based, post-hoc
Park et al. [52]	ICA	3 sequential CNNs	Vessel seg. + lumen contour + lesion match	Geometric, post-hoc
Liu et al. [8]	ICA	BasNet + PSPNet (sequential)	Vessel seg. + centreline + diameter fitting	Geometric, post-hoc
Huang et al. [28]	ICA	SAM + VM-UNet	Vessel seg. + dynamic-cohort stenosis det.	Centreline-based, post-hoc
This work (GCT-Net)	ICA	MaxViT + 2 U-Net + ROI head	Vessel seg. + stenosis seg. + QCA regression (joint, end-to-end)	Learned regression head with conformal prediction intervals

2.4 Comparative Analysis

Table 2.6 provides a unified summary of the reviewed methods across the three task categories. CNN-based and hybrid architectures are prevalent in pixel-level vessel segmentation, while transformer-based architectures lead on ECG-based lesion classification. Quantification methods build on segmentation outputs and report RMSE values that are competitive with manual stenosis grading.

Table 2.6: Comparison of Deep Learning Methods for Cardiovascular Disease Detection.

Study	**Architecture**	**Dataset**	**Result**	**Task**
Vessel Segmentation				
Liu et al. [47]	YOLOv8 Ensemble	ARCADE	F1 = 0.429	Vessel Segmentation
Liu et al. [8]	BasNet + PSPNet	Fuwai	F1 = 0.879	Vessel Segmentation
Alirr et al. [45]	ResUNet	ASOCA	DSC = 0.867	Coronary Segmentation
Zhao et al. (2020) [46]	MIMS U-Net	294 ICAs	DSC = 0.833	Vessel Seg. + Stenosis
Zhao et al. (2021) [9]	FP-U-Net++	314 ICAs	DSC = 0.890	Vessel Seg. + Stenosis
Lesion Detection				
Abrar et al. [42]	Swin Transformer	Kaggle ECG	Acc = 0.998	ECG Classification
Pokhrel et al. [49]	ConvNeXt-V2 + Mask R-CNN	ARCADE	F1 = 0.535	Stenosis Detection
Ansari [50]	Grounding DINO	ARCADE	mAP@0.50 = 0.259	Stenosis Detection
Li et al. [7]	LT-YOLO	CADICA	mAP@0.50 = 0.764	Video Stenosis
Wang et al. [43]	Faster R-CNN + ResNet	CCTA	AUC = 0.89	Calcified Plaque
Gupta et al. [44]	EfficientNet	CCTA	AUC = 0.95	Stenosis Detection
Lesion/Stenosis Quantification				
Liu et al. [8]	BasNet + PSPNet	Fuwai	RMSE = 0.064	QCA (LL, MLD, RVD, DS)
Zhao et al. (2020) [46]	MIMS U-Net	294 ICAs	TPR = 0.667	Stenosis % (4 grades)
Zhao et al. (2021) [9]	FP-U-Net++	314 ICAs	TPR = 0.684	Stenosis % (4 grades)

2.5 Conclusion

This chapter has surveyed deep learning approaches for cardiovascular disease detection in three groups — vessel segmentation, lesion detection, and stenosis quantification — and additionally reviewed multi-task and joint-quantification methods related to the present work (Section 2.3.4). Among single-task families, CNN-based approaches (YOLOv8, ResUNet, MIMS U-Net, FP-U-Net++) report the strongest vessel segmentation results, transformer-based architectures (Swin Transformer, DINO-DETR) lead on ECG-based lesion classification, and CNN-hybrid methods (Mask R-CNN, EfficientNet) remain competitive for angiographic lesion detection. Multi-stage training strategies, illustrated by Zhao et al. [9, 46], indicate that iterative refinement from coarse to fine segmentation combined with feature pyramid supervision improves both vessel segmentation accuracy and downstream stenosis detection and quantification.

Several common limitations emerge. Most studies are validated on a single centre, are restricted to single-view single-frame analysis, and rely on relatively small annotated datasets. Beyond these data-related issues, the review reveals a more specific methodological gap. Multi-task formulations have begun to appear in coronary CT angiography — including the joint wall/lumen-plaque segmentation of Lin et al. [22] and the joint segmentation–assessment design of MGFA-Net [21] — and earlier multi-task work on 3D vessel volumes [23] established the value of hard parameter sharing for vascular analysis. However, on X-ray non-invasive coronary angiography (ICA), the dominant quantification systems — AI-QCA [8], MPXA-1000 [52], MIMS U-Net [46], FP-U-Net++ [9], and SAM-VMNet [28] — follow a common template in which a segmentation network is trained first, followed by a separate post-processing algorithm that extracts centrelines, computes diameters, and derives stenosis percentages. This sequential design has two structural drawbacks. First, errors produced by the segmentation stage propagate unchecked into the quantification stage, since the two are optimised independently with different objectives. Second, the anatomical dependency between the tasks — a stenosis only exists on a vessel, and QCA quantifies a stenosis — is not exploited as a learning signal during training. From the multi-task learning perspective (Section 1.6.4), this represents a missed opportunity for positive task transfer and shared regularisation, particularly in single-centre settings where labels are scarce. To the best of our knowledge, no prior work jointly optimises vessel segmentation, stenosis localisation, and continuous QCA regression (%DS and lesion length) within a single end-to-end network on ICA imagery.

This observation motivates the contribution of the next chapter: a multi-task architecture (GCT-Net) in which a single shared CNN-Transformer encoder feeds three task-specific heads — vessel segmentation, stenosis localisation, and QCA regression — trained jointly on the Oasis Clinic dataset, with the QCA outputs additionally wrapped in a calibrated conformal predictor.

Chapter 3

Proposed Approach

3.1 Introduction

This chapter presents the methodology and experimental results of this work, structured as a three-stage research progression on the automated analysis of coronary artery disease. The progression covers three clinical sub-tasks — vessel segmentation, stenosis localisation, and Quantitative Coronary Analysis (QCA) regression — and three data modalities: X-ray non-invasive coronary angiography (ICA), the resting electrocardiogram (ECG), and the structured procedural report. Each stage was motivated by the limitations of the previous one, and the three are referred to throughout as the three *flows*:

1. *Flow 1 — Sequential pipeline.* Three independently trained models are chained: a vessel segmenter, a stenosis detector, and a QCA stage in which %DS and lesion length are derived from the segmentation outputs through geometric post-processing. Each stage is optimised against its own objective without any feedback from the downstream quantitative target. This flow reproduces the dominant ICA-quantification paradigm [8, 9, 28, 52] and quantifies its central weakness, the propagation of errors across stages (Section 3.5).

2. *Flow 2 — Single-source, multi-task.* A shared hybrid CNN-Transformer encoder feeds three task-specific heads that are predicted jointly in a single forward pass: two parallel segmentation decoders for the vessel and stenosis maps, and a learned QCA regression head conditioned on the predicted stenosis map through a lesion-aware ROI pooling mechanism. A stop-gradient decouples the localisation map from the regression objective, and the task losses are balanced automatically by homoscedastic uncertainty weighting (Sections 3.6 and 3.10).

3. *Flow 3 — Multi-source, multi-task.* The single-source model is extended along two complementary axes. First, two non-imaging modalities — a tabular resting-ECG feature vector and the structured fields parsed from the procedural report — are

fused into the visual bottleneck through Feature-wise Linear Modulation (FiLM) and supervised by four additional report-derived heads. Second, the QCA outputs are wrapped in a split-conformal predictor that attaches a distribution-free 90 % interval to each estimate (Section 3.11).

The three flows are positioned with respect to the prior art surveyed in Chapter 2 as follows. Flow 1 instantiates the dominant ICA-quantification paradigm — a sequential pipeline of independently trained models with post-hoc geometric quantification [8, 9, 28, 52] — and serves as the baseline against which the cost of the cascade is measured. Flow 2 introduces a unified multi-task design that combines four established lines of research: hard parameter sharing for joint medical segmentation–regression objectives [23, 25, 26], hybrid CNN-Transformer backbones [15, 27], homoscedastic uncertainty weighting for automatic loss balancing [16], and topology-preserving losses for tubular anatomies [55]. Flow 3 extends this design along two further axes that are unexplored on ICA: multi-task conditioning through FiLM fusion [34] and distribution-free interval estimation through split-conformal prediction [56], whose $1 - \alpha$ marginal coverage guarantee holds under the exchangeability of the calibration and test patients without parametric assumptions on the residuals and is verified empirically in Section 3.11.7.

Within this framing, the chapter makes five contributions: (i) a lesion-aware ROI pooling head with a stop-gradient and homoscedastic uncertainty weighting, enabling end-to-end multi-task learning of vessel segmentation, stenosis localisation, and QCA regression under partial supervision (Flow 2, Section 3.6); (ii) a FiLM-based multi-task fusion of ECG and procedural-report fields with four report-derived auxiliary heads (Flow 3, Section 3.11); (iii) a split-conformal predictor providing distribution-free 90 % intervals on the QCA outputs (Flow 3, Section 3.11.7); (iv) a comparative analysis of the three flows on the Oasis Clinic cohort and an external validation of the stenosis branch on the independent ARCADE benchmark (Sections 3.12 and 3.13); (v) the deployment of the trained system in a clinical web application (Section 3.14).

The chapter first describes the multi-source clinical dataset collected at Oasis Clinic (Ghardaïa, Algeria) and the unified preprocessing pipeline. It then presents the three flows in order, followed by a comparative analysis of the three, an external validation of the stenosis branch on the independent ARCADE benchmark, the deployment of the trained system in a clinical web application, and a discussion. All training experiments were conducted on the Kaggle platform using a single NVIDIA GPU (Tesla T4 for Flow 2, Tesla P100 for Flow 1 and Flow 3); the implementation is in PyTorch.

3.2 Dataset Description

The dataset used in this work was collected on-site from the catheterisation laboratory of Oasis Clinic, Ghardaïa, Algeria, over a one-month engagement, and comprises multi-source records from 200 patients with suspected Coronary Artery Disease (CAD). For each patient, three complementary sources were exported from the clinic PACS server under a secure DICOM-export protocol: the X-ray non-invasive coronary angiography (ICA) cine sequences in DICOM format, the resting electrocardiogram (ECG) as a PDF, and the procedural report as a Word (`.docx`) document.

Each patient is identified by a unique `PA<id>` folder. The raw angiographic archive consists of several multi-frame DICOM files (`IM0`, `IM1`, ...), each storing one acquired cine sequence (left or right coronary artery at different angulations) of 15 to 183 greyscale frames at a native resolution of 512512 pixels. The number of DICOM files per patient varies between 4 and 14. The DICOM header encodes the metadata relevant to this work — patient and study identifiers, image geometry, the pixel spacing required to convert pixel measurements to millimetres, the transfer syntax (JPEG 2000 lossless), and the vendor private tags written by the angiography console.

Beyond the raw cine data, two clinically prepared frames accompany each study, produced during the original cath-lab workflow:

- *Chosen frame*: a single representative DICOM frame selected manually by clinical staff at peak contrast opacification. It is retained only as a cross-reference for validating the automatic frame selector (Section 3.6.3); the proposed pipeline does not depend on it (Figure 3.1).
- *Annotated frame*: the same frame after annotation by the cath-lab QCA software, in which the stenotic segment is highlighted and the QCA measurements — the reference vessel diameter, the minimal lumen diameter, and the lesion length — are written into the DICOM private tags (Figure 3.2).

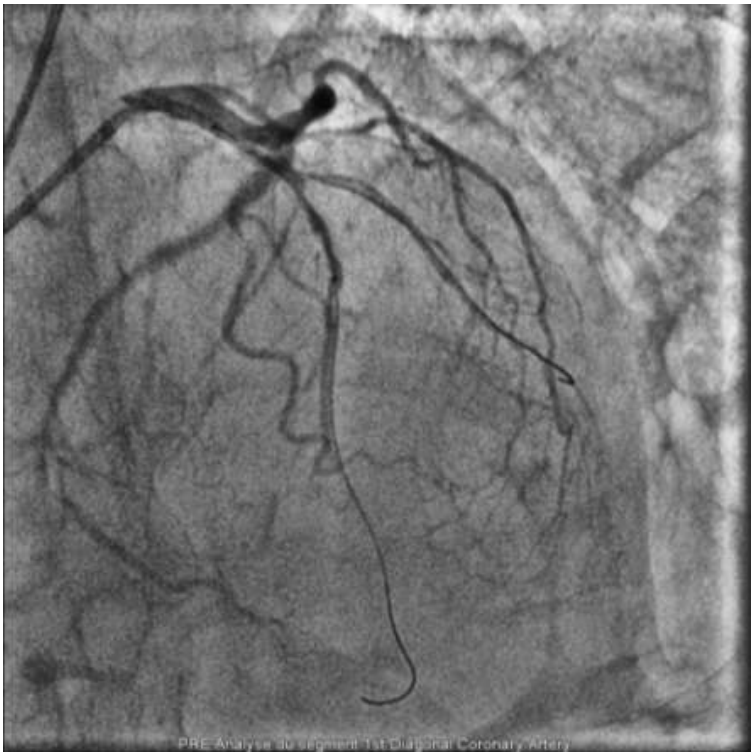


Figure 3.1: The chosen frame for a representative patient: a single best-quality frame selected manually by clinical staff at peak contrast opacification, with minimal motion blur and clear vessel visibility.

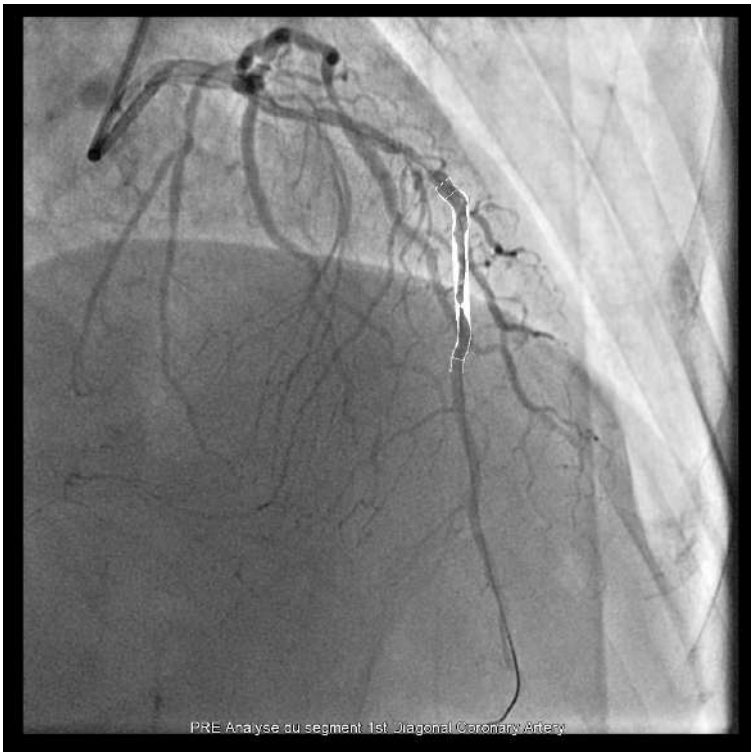


Figure 3.2: The annotated frame for the same patient, produced by the cath-lab QCA software: the stenotic segment is highlighted and the QCA measurements (reference vessel diameter, minimal lumen diameter, and lesion length) are written into the DICOM private tags.

3.3 Preprocessing and Ground-Truth Creation

The raw multi-source archive is transformed into a clean training set through a single deterministic pipeline. This section describes the image preprocessing applied to the angiographic frame, the manual classification and annotation that produce the training labels, the conversion of the two non-imaging modalities into model-ready feature vectors, and the final assembly of the paired dataset.

3.3.1 Image Preprocessing

For each patient, the angiographic frame is read with the `pydicom` library [12] (with a GDCM backend for JPEG 2000 decompression), after which the following steps are applied in order:

1. Min–max intensity normalisation, $I_{\text{norm}} = (I - I_{\min})/(I_{\max} - I_{\min} + \epsilon)$ with $\epsilon = 10^{-8}$, mapping the raw DICOM intensities to $[0, 1]$ before rescaling to `uint8`; this standardises the intensity range across studies acquired with different exposure settings.

2. CLAHE contrast enhancement [11] with clip limit 2.0 and tile grid 88, which improves local vessel visibility while avoiding noise amplification in homogeneous background regions.

3. Mild Gaussian denoising with a 33 kernel, attenuating residual sensor noise while preserving vessel edges.

4. Bilinear resizing to 512512, matching the input resolution at which the MaxViT backbone was pretrained.

5. Channel replication and ImageNet normalisation: the single greyscale channel is replicated to three channels and normalised with the ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), so that the ImageNet-pretrained MaxViT encoder receives inputs within its training distribution.

3.3.2 Frame Sorting

A single DICOM run contains dozens of frames spanning the full contrast cycle, of which only a narrow window is diagnostically usable. To build a supervisory signal for frame quality, each frame of a curated subset of runs was manually classified into two categories: *pre-contrast*, in which the vessel tree is not yet opacified and the lumen is unclear, and *peak-contrast*, in which the vessel is fully opacified and optimal for analysis. The classification was performed manually to maximise label accuracy. This sorted set provides the ground truth that underpins the automatic best-frame selection module (Section 3.6.3), which later replaces the manual selection step entirely at deployment time.

3.3.3 Segmentation Ground-Truth Creation

Two binary pixel-level masks were created manually on a representative subset of patients using the open-source GIMP raster editor, as no pixel-level annotations were provided by the clinic.

Binary vessel mask (`vessel_gt.png`). A pixel-level annotation of the entire coronary tree, in which white pixels (255) denote vessel tissue and black pixels (0) denote background. This mask supervises the vessel-segmentation task (Figure 3.3).

Stenosis mask (`<id>_stenosis_gt.png`). A pixel-level binary mask marking the diseased vessel region, the most challenging artefact to produce, in which white pixels denote the stenotic segment and black pixels denote non-stenotic tissue. This mask supervises the stenosis-localisation task (Figure 3.4).

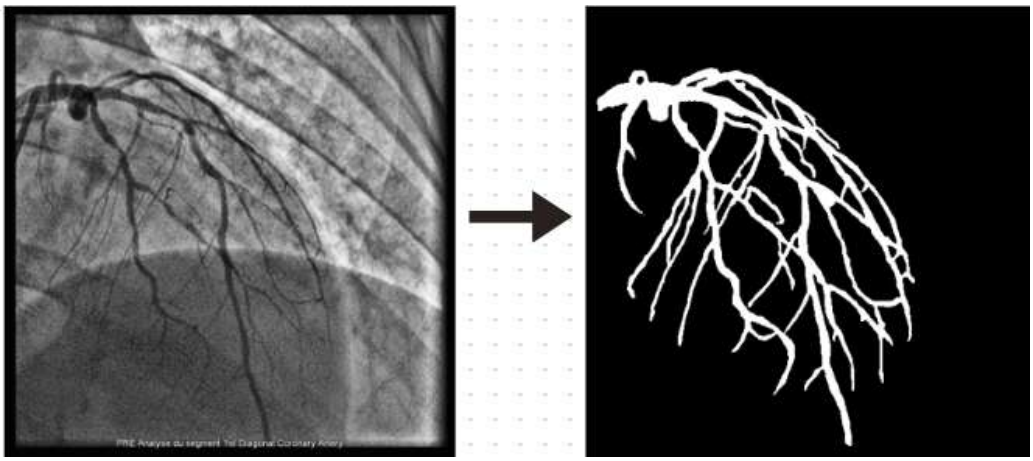


Figure 3.3: Binary vessel mask creation. *Left*: the chosen frame. *Right*: the manually created binary vessel mask (`vessel_gt.png`), in which white pixels denote vessel tissue and black pixels denote background. The mask delineates the full coronary tree and was drawn in GIMP on a representative subset of patients.

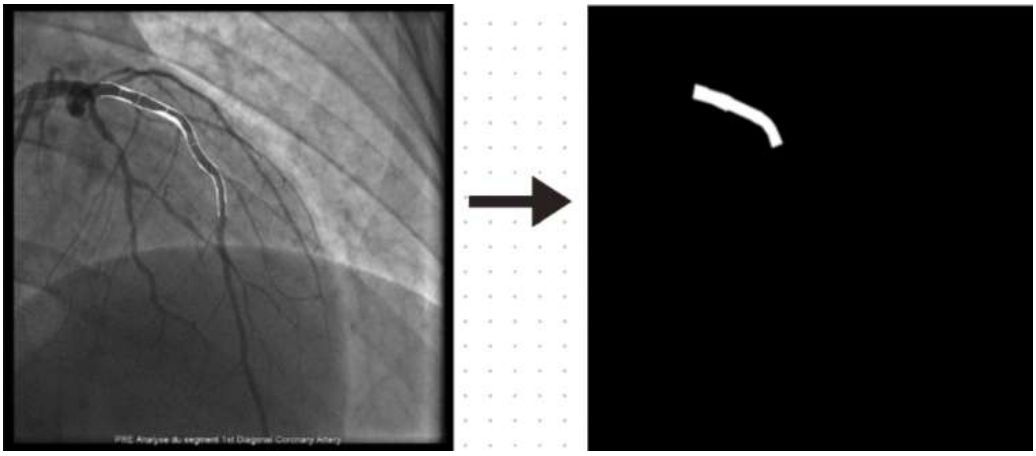


Figure 3.4: Binary stenosis mask creation. *Left*: the annotated frame. *Right*: the manually created binary stenosis mask (`<id>_stenosis_gt.png`), in which white pixels mark the diseased vessel segment and black pixels mark non-stenotic tissue. This is the most challenging artefact to produce and was drawn in GIMP on a representative subset of patients.

During preprocessing these masks undergo the same geometric resizing as the frame, using nearest-neighbour interpolation to preserve their binary values; the soft pseudo-labels produced by the U-Net teacher (Section 3.3.6) are instead resized with bilinear interpolation. Non-empty stenosis masks are dilated with a 55 elliptical kernel (two iterations) to compensate for the thinness of the manual contours. During training, the `Albumentations` library [20] applies a synchronised data-augmentation pipeline (Section 3.8) to the frame and both masks simultaneously, preserving pixel-wise alignment.

3.3.4 Non-Imaging Feature Extraction

The two non-imaging modalities are converted offline into compact, fixed-length numeric records before training, so that no raw PDF or document parsing occurs inside the training loop.

ECG features (`Clinical12_Features.csv`)

Although the resting 12-lead ECG is originally exported as a PDF tracing, it enters the model not as an image but as a compact numerical vector. The PDF is processed by the open-source NeuroKit2 toolkit [4] — a pre-trained, clinically validated and reproducible ECG analysis library (D. Makowski) — which is run once per patient in inference mode. Each tracing yields twelve standardised clinical measurements plus an extraction-confidence score, forming a 13-dimensional feature vector that is joined to the patient record (Table 3.1). Using a pre-trained extractor removes the slow, error-prone manual digitisation of the tracing and yields consistent, standardised features across all cases. The extraction-confidence value additionally serves as a quality flag that gates the corre-

sponding auxiliary losses when the ECG is missing or unreliable; the use of this vector as a conditioning signal is detailed in Flow 3 (Section 3.11).

Table 3.1: The twelve ECG features extracted by NeuroKit2 from each 12-lead tracing, together with the extraction-confidence score. Binary flags are encoded as 0/1; the confidence score lies in $[0, 1]$.

Token	**Feature**	**Type / unit**	**Example**
HR	Heart Rate	bpm	72
PR	PR Interval	ms	160
QRS	QRS Duration	ms	88
QT	QT Interval	ms	400
QTc	Corrected QT (Bazett)	ms	430
AXIS	QRS Electrical Axis	degrees	+42
ST_CHG	ST-Segment Change	mm	−1.2
TWAVE	T-Wave Inversion	0/1	1
RBBB	Right Bundle Branch Block	0/1	0
LBBB	Left Bundle Branch Block	0/1	0
AFIB	Atrial Fibrillation	0/1	0
CONF	Extraction Confidence	$[0, 1]$	0.93

Report features (`report_features.csv`)

The procedural report is a free-text narrative written by the cardiologist after the percutaneous coronary intervention. In raw form it is unusable as a training label, so it is parsed once by a large-language-model extractor (Gemini 3.1 Pro) into a structured table of typed fields, consumed as feature fields rather than as raw text (Table 3.2). Negation and uncertainty are detected explicitly during parsing, and a confidence score is attached to each extracted entity. Each extracted report then passes a validation step to ensure the language model has not introduced errors. The extractor converts each free-text report into a fixed-length numeric vector spanning the procedural indication, the treated vessel, the stent geometry, the procedural flow grade, the outcome, and a set of binary complication flags. These parsed fields supply the report-derived supervision heads introduced in Flow 3 (Section 3.11).

Table 3.2: The structured tokens extracted from each free-text procedural report by the large-language-model parser. Categorical fields are later one-hot encoded, numeric fields normalised, and the treated-vessel field treated as a multi-label over the major coronary arteries.

Token	**Field**	**Type**	**Example**
INDICA	Indication (type of ACS triggering the procedure)	Category	NSTEMI
ARTERY	Treated vessel (PCI target territory)	Multi-label	IVA prox.
STNT_D	Stent diameter (nominal)	mm	2.75
STNT_L	Stent length	mm	28
N_STNT	Number of stents deployed	Integer	1
TIMI	TIMI flow grade (0=no flow, 3=normal)	0–3	3
OUTCOM	Procedural outcome	Category	SUCCESS
COMPL	Complication reported	0/1	0
DISECT	Coronary dissection post-stenting	0/1	0
PERFOR	Vessel perforation during procedure	0/1	0
NOBLOW	No-reflow event (microvascular obstruction)	0/1	0
CONFID	Extraction confidence	$[0, 1]$	0.92

3.3.5 Per-Patient Dataset Organisation

The extracted and processed artefacts are stored under a per-patient directory structure, `DICOM/PA<id>/ST0/`, that co-locates the raw imaging, the derived annotation artefacts, and the structured non-imaging metadata, so that all modalities for a patient remain aligned under a single identifier. The layout is shown in Figure 3.5.

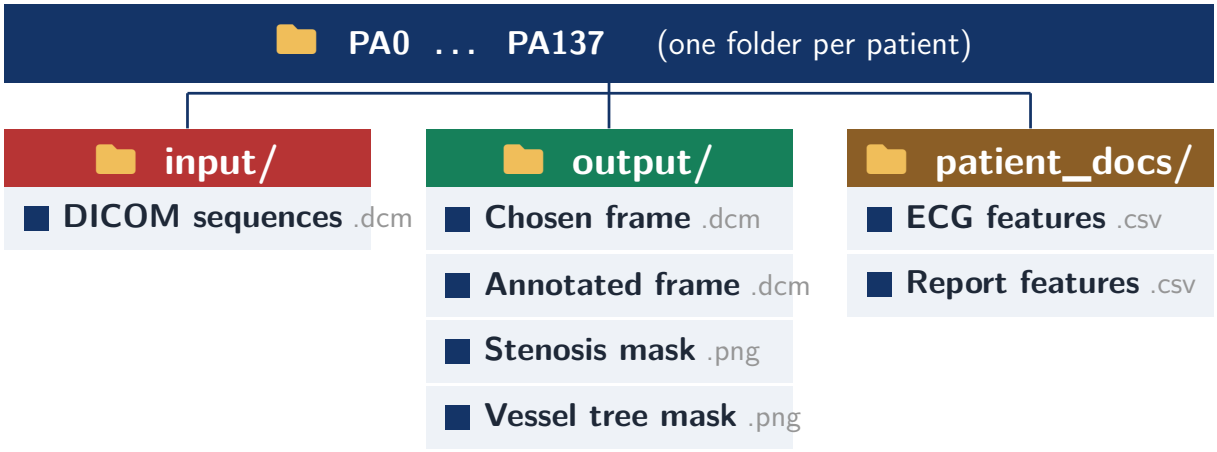


Figure 3.5: Per-patient dataset folder structure. Each patient occupies one `PA<id>` folder containing three sub-directories: `input/` holds the raw multi-frame DICOM cine sequences; `output/` holds the clinically chosen and annotated frames and the manually created vessel and stenosis masks; and `patient_docs/` holds the structured non-imaging metadata (the tabular ECG features and the parsed report fields). Co-locating all modalities under one identifier keeps them aligned for multi-source training.

3.3.6 Multi-Source Ground Truth

Because no single label source covers all 200 patients, the proposed system is trained from heterogeneous sources of supervision. The imaging tasks of Flow 1 and Flow 2 draw on the four signals summarised in Table 3.3, while the multi-source extension of Flow 3 additionally uses the ECG and report-derived features (Section 3.11).

Table 3.3: Multi-source ground-truth signals used to train GCT-Net.

Source	Type	Used for	Validity flag
Manual vessel mask (`vessel_gt.png`)	Hard binary mask	Vessel head	`vessel_is_hard` = 1.0
U-Net teacher pseudo-label	Soft probability map (distillation)	Vessel head	`vessel_is_hard` = 0.5
Manual stenosis mask (`<id>_stenosis_gt.png`)	Hard binary mask	Stenosis head	`stenosis_valid` = 1.0
DICOM private tags (0009,1013), (0009,1014), (0009,1016)	Numeric values	QCA head	`qca_valid` = 1.0

Manual vessel masks are the cleanest signal but are available only for a subset of

patients. Where they are missing, the system falls back to soft pseudo-labels produced by an existing U-Net teacher model — a form of knowledge distillation (Section 1.6.6). These contributions are down-weighted by a factor of 0.5 relative to manual masks, reflecting their lower confidence. Manual stenosis masks were drawn by clinical staff for the patients with diagnostically significant lesions. The QCA targets — the stenosis percentage $\%DS$ and the lesion length ℓ in millimetres — are extracted directly from three private DICOM tags written by the angiography console. They are defined by the two separate relations

$$\%DS \;=\; \left(1 - \frac{\mathrm{MLD}}{\mathrm{RVD}}\right) 100, \tag{3.1}$$

$$\ell \;=\; \mathrm{LesionLength}_{\mathrm{mm}}, \tag{3.2}$$

where Equation (3.1) gives the percentage diameter stenosis and Equation (3.2) gives the lesion length; the two quantities are read independently from the console and ℓ does not enter the $\%DS$ formula. Here MLD (tag `(0009,1014)`) is the minimal lumen diameter, RVD (tag `(0009,1013)`) is the reference vessel diameter, and the lesion length (tag `(0009,1016)`) is the longitudinal extent of the lesion. The extracted values are validated (positive RVD, non-negative MLD and length) and clipped to the physiologically meaningful range $\%DS \in [0, 100]$; tag-extraction failures fall back to `qca_valid` $= 0$. To avoid repeated DICOM reads during training, all QCA values are pre-extracted once into a JSON cache.

The signals coexist in the same dataset under partial supervision: each patient sample carries a binary validity flag for each task, and the multi-task loss masks out invalid contributions per sample (Section 3.7). This is the practical mechanism that makes a unified multi-task model trainable on heterogeneous, only partially overlapping label sources.

3.4 Methodology: The Three Architectural Flows

Having described the multi-source dataset and the unified preprocessing pipeline, the following sections present the three flows in order of increasing integration. Each is developed as a self-contained design with its own architecture, training procedure, and experimental results, and each is motivated by the limitations of the flow that precedes it. Flow 1 establishes the sequential baseline; Flow 2 unifies the three imaging tasks behind a shared encoder; and Flow 3 extends that encoder with the non-imaging modalities and calibrated uncertainty estimation.

3.5 Flow 1 — Sequential Pipeline

The first flow reproduces the prevailing ICA-quantification paradigm, in which the problem is decomposed into three separate models, each trained and evaluated in isolation and then chained at inference time so that the output of one stage becomes the input of the next. The pipeline comprises a vessel-segmentation stage, a stenosis-localisation stage, and a QCA quantification stage, each with its own architecture, loss function, and training schedule, and with no gradient flowing between them.

This decoupled construction is the natural starting point for the work. It is straightforward to build and validate one stage at a time, it provides a faithful baseline against which the unified models are measured, and, most importantly, it exposes the structural weakness of the paradigm: an error introduced at one stage propagates uncorrected to the next, since no stage is optimised against the downstream quantitative target. This cascading-error limitation is precisely what motivates the jointly optimised single-network designs of Flow 2 and Flow 3. The overall architecture is shown in Figure 3.6.

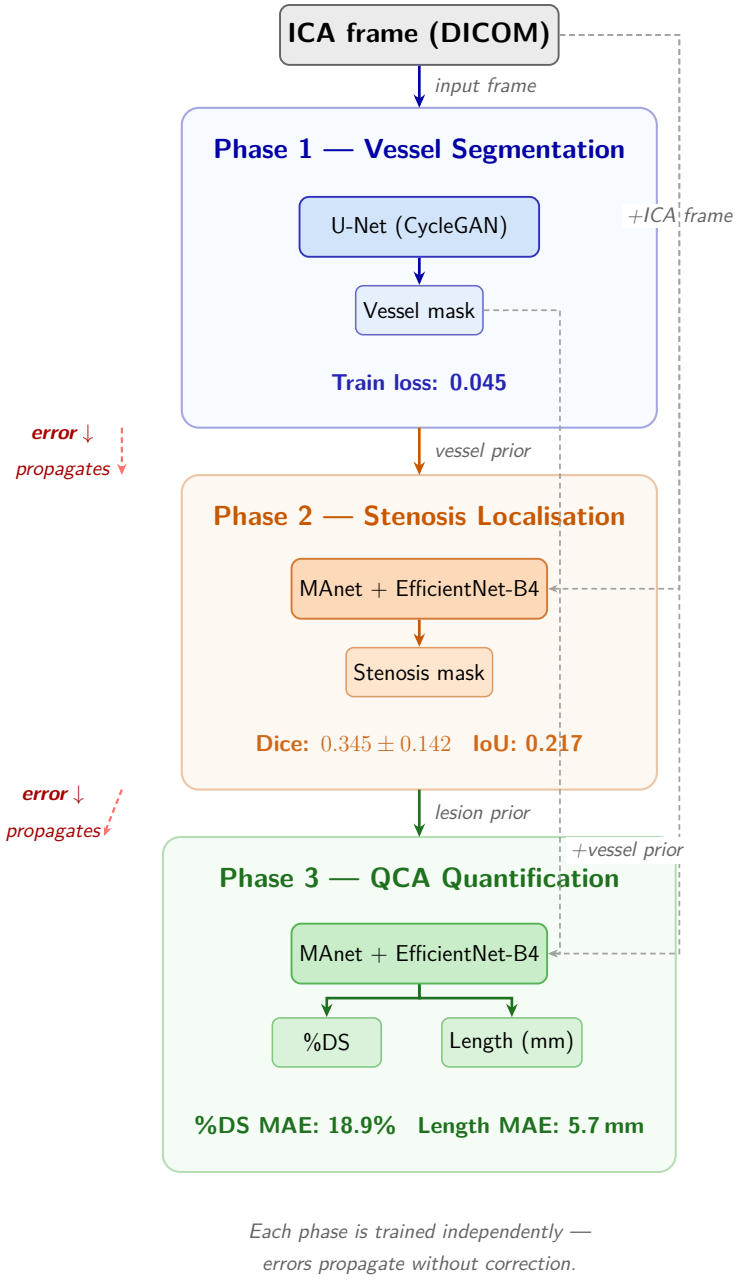


Figure 3.6: Architecture of the Flow 1 sequential pipeline (three phases, vertical layout). Phase 1 (blue): U-Net vessel segmenter. Phase 2 (orange): MAnet+EfficientNet-B4 stenosis localiser. Phase 3 (green): MAnet+EfficientNet-B4 QCA model with a regression head. Solid arrows carry the primary stage output; dashed arrows (right) show auxiliary inputs from earlier stages. Red arrows (left) highlight the cascading-error coupling.

Phase 1 — Vessel Segmentation

The vessel-segmentation stage has a dual role: it produces a binary mask of the full coronary tree for the current patient, and this mask then serves as the anatomical prior that constrains both the stenosis-localisation and the QCA stages that follow. Because no coronary-artery segmentation model existed for the Oasis Clinic acquisition protocol, the stage had to be built from scratch. The development proceeded through three successive

attempts, each one exposing a specific failure mode of the previous design and addressing it with a targeted architectural or supervisory change.

Attempt 1 — Naive U-Net (Single-Pass). The first attempt trained a standard U-Net [5] to predict vessel pixels directly from a single contrast-phase angiographic frame using a `BCEWithLogitsLoss`, the standard pixel-wise binary cross-entropy applied to the sigmoid of the network logits z_i:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\, y_i \log \sigma(z_i) + (1 - y_i) \log\big(1 - \sigma(z_i)\big)\Big], \tag{3.3}$$

where z_i is the predicted logit at pixel i, $\sigma(\cdot)$ the sigmoid so that $\hat{y}_i = \sigma(z_i)$ is the predicted vessel probability, $y_i \in \{0, 1\}$ the ground-truth label, and N the total number of pixels. The input was the chosen frame; no preprocessing or anatomical prior was provided. The network produced scattered, noisy prediction fragments with no anatomical coherence. The root cause is that the contrast-filled coronary arteries share a similar local intensity with the static background structures present in every frame — the bony ribcage, the thoracic spine, and the catheter guide-wire — so a pixel-level loss with no spatial context gives the network no incentive to distinguish vessel from non-vessel regions that happen to have the same grey value. The guide-wire in particular, being a thin high-contrast wire running across the image, was consistently included in the predicted mask. This attempt established the baseline failure mode: without background modelling, single-frame vessel segmentation on raw angiograms is not viable.

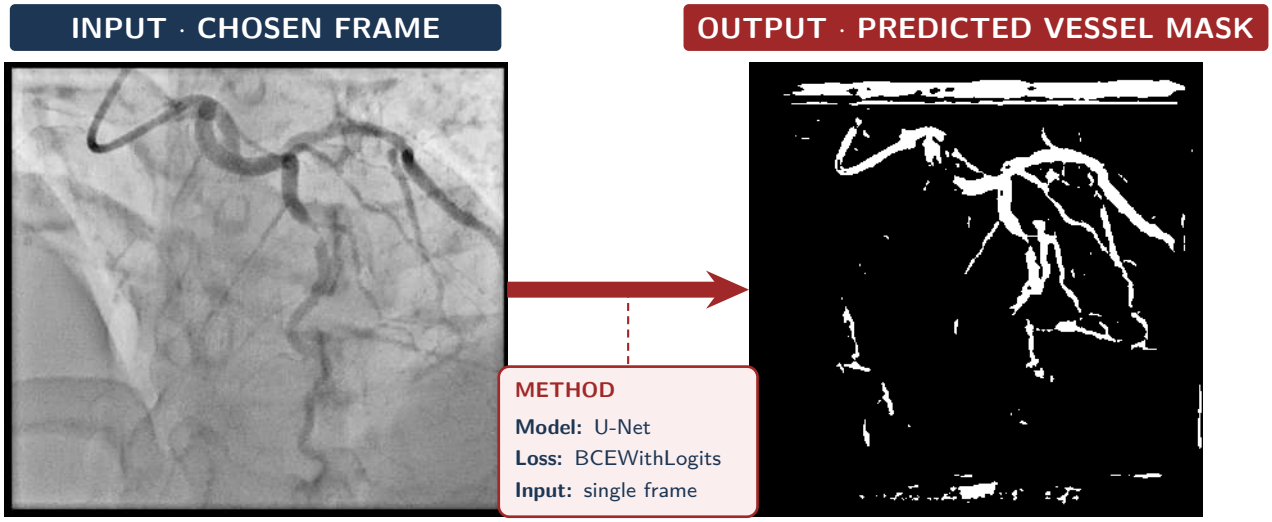


Figure 3.7: Vessel segmentation Attempt 1. The raw chosen frame (left) contains vessels, bones, and the guide-wire at similar intensity. The naive U-Net output (right) shows scattered, noisy fragments with no anatomical coherence: the network cannot distinguish vessels from the bone and guide-wire background, which share the same local grey values.

Attempt 2 — Two-Pass Background Suppression. The second attempt introduced a two-pass training scheme designed to explicitly model and remove the static background before attempting vessel segmentation. The scheme is built directly on the frame-sorting step of the preprocessing pipeline (Section 3.3.2), in which every frame of a DICOM run is classified as either *pre-contrast* — acquired before the iodinated contrast agent reaches the coronary tree, so no vessel signal is present — or *peak-contrast* — fully opacified and optimal for analysis. This sorted partition is what makes the two passes possible. In the first pass, the network is trained on the pre-contrast frames to learn a model of the static background (the bony ribcage, the thoracic spine, and the catheter guide-wire), which are the only structures visible at that stage. In the second pass, the learned background representation is used as a reference that is subtracted from the corresponding peak-contrast frame, so the subsequent segmentation head sees a signal from which the static structures have already been suppressed. This approach successfully recovered a coherent coronary-tree mask with an anatomically plausible structure, and the majority of off-vessel activations disappeared. However, the guide-wire and a handful of residual bone edges remained in the prediction: the background model was not perfectly accurate at those locations, and the residual contrast they created was still interpreted as vessel signal. While a major improvement over Attempt 1, the output was not yet clean enough to serve as a reliable anatomical prior for the downstream stages.

Loss function (Attempt 2). The two passes are trained with a different objective each. In **Pass 1**, the network learns to *reconstruct* the static background from the pre-contrast frames, in which no vessel signal is present; writing x^{pre} for the pre-contrast frame and $\hat{b} = \mathcal{F}_\theta(x^{\text{pre}})$ for the background image predicted by the reconstruction network $\mathcal{F}_\theta$ with parameters θ, the reconstruction objective is the pixel-wise mean squared error

$$\mathcal{L}_{\text{bg}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{b}_i - x_i^{\text{pre}}\right)^2, \tag{3.4}$$

where $\hat{b}_i$ and x_i^{pre} are the predicted and observed background intensities at pixel i, and N is the total number of pixels (as in Equation (3.3)). In **Pass 2**, the learned background is subtracted from the corresponding peak-contrast frame to form the suppressed input $\tilde{x} = x^{\text{peak}} - \hat{b}$, and the vessel head is trained on $\tilde{x}$ with the same pixel-wise binary cross-entropy of Equation (3.3). The two passes are optimised sequentially: the background model of Equation (3.4) is trained to convergence and then frozen, so the only supervisory signal seen by the vessel head is the background-suppressed cross-entropy.

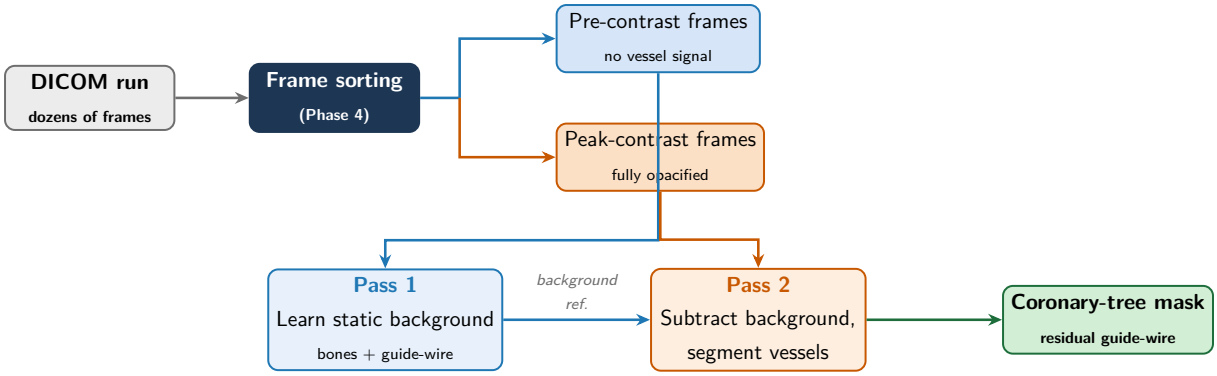


Figure 3.8: The two-pass background-suppression scheme of Attempt 2, driven by the Phase 4 frame-sorting step. Every frame of the DICOM run is first sorted into pre-contrast and peak-contrast classes. Pass 1 (blue) learns the static background — bones and guide-wire — from the pre-contrast frames, where no vessel signal is present, under the reconstruction loss of Equation (3.4). Pass 2 (orange) subtracts this learned background reference from the corresponding peak-contrast frame and segments the vessels from the suppressed signal with the cross-entropy of Equation (3.3), yielding a coherent coronary-tree mask in which a faint guide-wire residual still leaks through.

Attempt 3 — CycleGAN Pretraining + Supervised Fine-Tuning (Final Model). The third and final attempt retained the two-pass suppression strategy of Attempt 2 and added two further components to eliminate the remaining artefacts: unsupervised domain pretraining via CycleGAN [13], followed by supervised fine-tuning on the Oasis Clinic dataset itself.

The CycleGAN pretraining phase adapts the encoder weights to the angiographic image domain before any supervised signal is applied. A CycleGAN is trained to learn an unpaired image-to-image translation between the raw angiographic frames and a target domain of cleaner vessel images, forcing the encoder to extract domain-relevant features rather than random texture statistics. The pretrained weights are then used to initialise the U-Net encoder for the subsequent supervised stage.

Loss function (CycleGAN pretraining). Let $G : \mathcal{X} \to \mathcal{Y}$ map a raw angiographic frame to the clean vessel domain and $F : \mathcal{Y} \to \mathcal{X}$ the inverse mapping, with D_X and D_Y the corresponding domain discriminators. Here $\mathcal{X}$ and $\mathcal{Y}$ denote the raw-angiogram and clean-vessel image domains with data distributions $p_{\mathcal{X}}$ and $p_{\mathcal{Y}}$, $\mathbb{E}[\cdot]$ is the expectation taken over those distributions, and $D_Y(\cdot) \in [0, 1]$ is the probability that its argument is a real image of domain $\mathcal{Y}$ (and symmetrically for D_X). Each mapping is trained against an adversarial objective

$$\mathcal{L}_{\mathrm{GAN}}(G, D_Y) = \mathbb{E}_{y \sim p_{\mathcal{Y}}}\big[\log D_Y(y)\big] + \mathbb{E}_{x \sim p_{\mathcal{X}}}\big[\log\big(1 - D_Y(G(x))\big)\big], \tag{3.5}$$

and the unpaired mappings are tied together by the cycle-consistency loss

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_x\big[\,\|F(G(x)) - x\|_1\,\big] + \mathbb{E}_y\big[\,\|G(F(y)) - y\|_1\,\big], \tag{3.6}$$

which forces a frame to be recovered after a round trip through both mappings ($x \to G(x) \to F(G(x)) \approx x$), where $\|\cdot\|_1$ is the pixel-wise ℓ_1 norm. The full pretraining objective is

$$\mathcal{L}_{\text{CycleGAN}} = \mathcal{L}_{\text{GAN}}(G, D_Y) + \mathcal{L}_{\text{GAN}}(F, D_X) + \lambda_{\text{cyc}}\,\mathcal{L}_{\text{cyc}}(G, F), \tag{3.7}$$

with the cycle weight set to $\lambda_{\text{cyc}} = 10$. Only the encoder weights of G are retained; they initialise the U-Net encoder for the supervised stage that follows.

The supervised fine-tuning stage trains the full U-Net on 70 samples drawn from the Oasis Clinic patient archive. For each patient, the chosen frame is processed and paired with the vessel mask produced by the model itself as a pseudo-label, formatted into an image–mask export. This self-supervised strategy lets the network refine its predictions on the exact acquisition domain it will be deployed on, without requiring any additional external data. The network architecture is a custom U-Net with feature width `dim` = 32 and bilinear upsampling ($n\_\text{channels} = 1$, $n\_\text{classes} = 1$). Training is run for 300 epochs with a batch size of 8, using the Adam optimiser at a learning rate of $\eta = 10^{-4}$ with a `ReduceLROnPlateau` scheduler (patience = 5, reduction factor 0.5). The loss function is the same pixel-wise binary cross-entropy of Equation (3.3), here a `BCEWithLogitsLoss` applied to the normalised (512512, single greyscale channel) frames. The training loss decayed monotonically from 0.637 at epoch 1 to a converged value of **0.045** at epoch 300, with the best checkpoint saved at each new minimum. This combination eliminated the guide-wire and bone-edge artefacts entirely and produced clean, anatomically faithful coronary-tree segmentations.

Because no manual vessel masks were withheld as a separate test partition at this stage of development, the final vessel model is assessed qualitatively: its per-patient predictions are visually inspected for anatomical coverage and exported as `prediction.png` files. The same trained model is re-used twice later in the unified pipeline: as the teacher network that drives automatic best-frame selection (Section 3.6.3), and as the source of soft pseudo-label supervision for the vessel head of the unified model (Section 3.3.6).

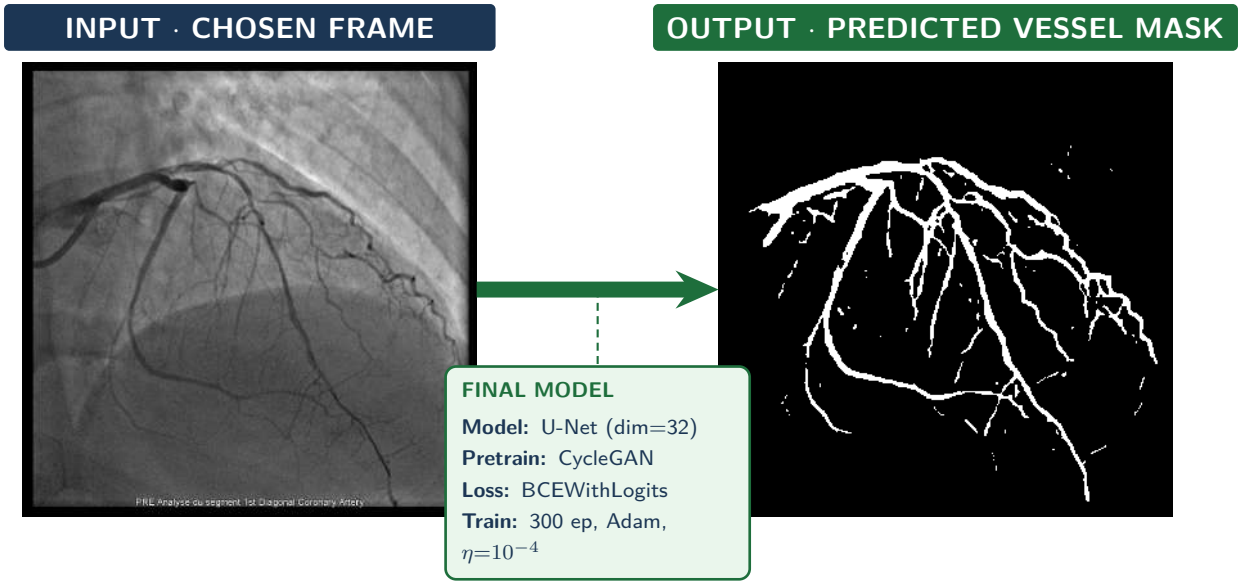


Figure 3.9: Vessel segmentation Attempt 3 (final model). The CycleGAN-pretrained U-Net fine-tuned on 70 Oasis Clinic patient frames (left: input frame; right: output mask) produces a clean binary segmentation of the full coronary tree. All guide-wire and bone-edge artefacts present in Attempts 1 and 2 are eliminated. Final training loss: ≈ 0.045 at epoch 300.

Phase 2 — Stenosis Localisation

The stenosis-localisation stage takes the DICOM frame and the vessel mask produced by Phase 1 as its inputs and outputs a binary lesion mask delineating the diseased segment of the coronary tree. Its purpose within the pipeline is twofold: it tells Phase 3 *where* to look for the lesion when extracting QCA measurements, and it provides a spatially precise mask against which clinical ground truth can be compared. Like the vessel stage, this stage was developed through three successive attempts of increasing supervision, each one diagnosing and correcting the specific failure mode of the previous design.

Attempt 1 — Naive U-Net. The first attempt was a plain U-Net trained directly on the annotated frame with no information about where the coronary tree lies. The loss was a combination of Dice loss and vessel-masked binary cross-entropy. The Dice loss directly optimises the overlap between prediction and ground truth,

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} \hat{y}_i\, y_i + \epsilon}{\sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} y_i + \epsilon}, \tag{3.8}$$

with $\hat{y}_i = \sigma(z_i)$ and ϵ a small smoothing constant, while the vessel-masked binary cross-entropy restricts the per-pixel BCE to the coronary region defined by the vessel mask

$m_i \in \{0, 1\}$,

$$\mathcal{L}_{\text{vBCE}} = \frac{\sum_{i=1}^{N} m_i \left[-y_i \log \sigma(z_i) - (1 - y_i) \log \left(1 - \sigma(z_i)\right)\right]}{\sum_{i=1}^{N} m_i + \epsilon'}, \tag{3.9}$$

so that pixels outside the vessel tree ($m_i = 0$) contribute no gradient ($\epsilon' = 10^{-8}$). The two terms are combined as an unweighted sum,

$$\mathcal{L}_{\text{sten}}^{(1,2)} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{vBCE}}, \tag{3.10}$$

which is the objective used identically in both Attempt 1 and the vessel-gated Attempt 2 below — the two attempts differ only in their input, not in their loss. The outcome was immediately poor: the network fired indiscriminately across the full image, producing numerous false-positive activations on bone shadows and on the catheter guide-wire, both of which locally mimic the contrast appearance of a stenotic lesion segment. Because the network had no coronary prior, it had no reason to confine its predictions to the vessel tree, and it learned instead to respond to any compact, moderately bright region in the frame — a pattern that matches the artefacts as well as the true lesions. This demonstrated that a stenosis detector without anatomical grounding is not feasible on raw angiographic frames.

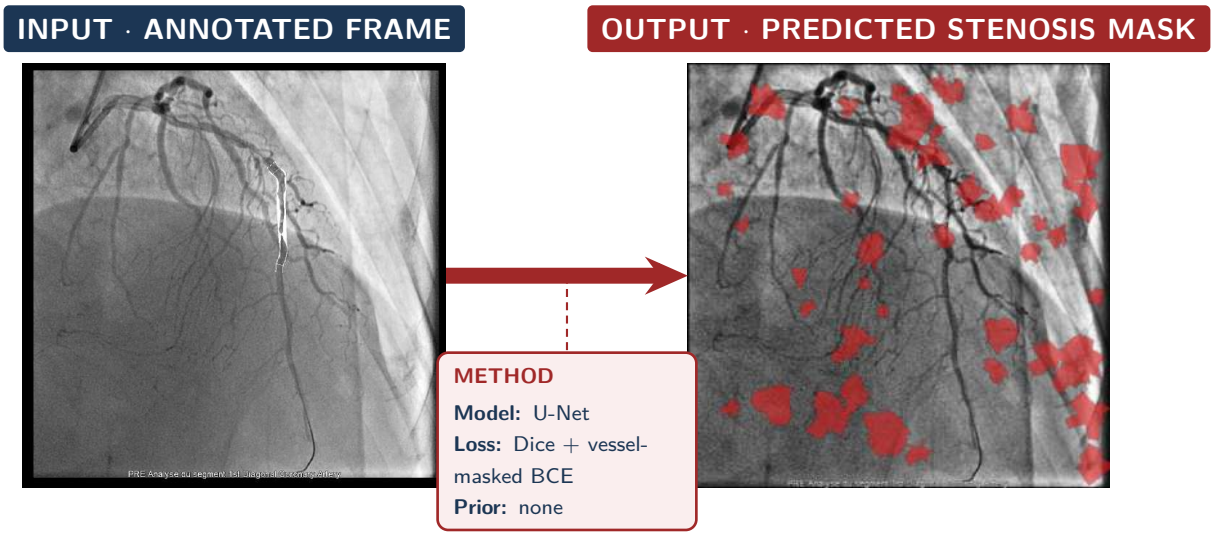


Figure 3.10: Stenosis localisation Attempt 1. The annotated frame (left) is fed to a naive U-Net with no coronary prior. The output (right) shows numerous false-positive activations (red) scattered indiscriminately across the full image, firing on bone shadows and the guide-wire which locally mimic the contrast appearance of a lesion. No anatomical constraint confines the predictions to the coronary tree.

Attempt 2 — Vessel-Gated Detector. The second attempt introduced the vessel mask from Phase 1 as an explicit anatomical prior. The pre-trained vessel-segmentation output and the annotated frame are stacked into a two-channel input tensor [angiogram, vessel mask]

of shape 2512512, so the detector is told at every pixel whether it lies inside the coronary tree or not. The architecture was upgraded to a **MAnet** decoder (Multi-scale Attention network) with an **EfficientNet-B4** encoder [3], pretrained on ImageNet. The loss remained the Dice-plus-vessel-masked-BCE sum of Equation (3.10) (Equations (3.8) and (3.9)).

This gating strategy was highly effective at removing off-vessel false positives: activations on bone shadows and the guide-wire disappeared almost entirely, and precision improved substantially. However, a second failure mode emerged: the predicted lesion mask was coarse and incomplete. The reason is that the only positive supervisory signal in Attempt 2 came from the annotated region visible in the frame, a small and locally weak signal that is insufficient for the network to recover the full spatial extent of the lesion boundary. The network learned to detect the approximate location of the lesion but could not recover its precise shape or extent, yielding masks suitable for rough localisation but not for the downstream QCA measurement, which requires accurate lesion-boundary delineation.

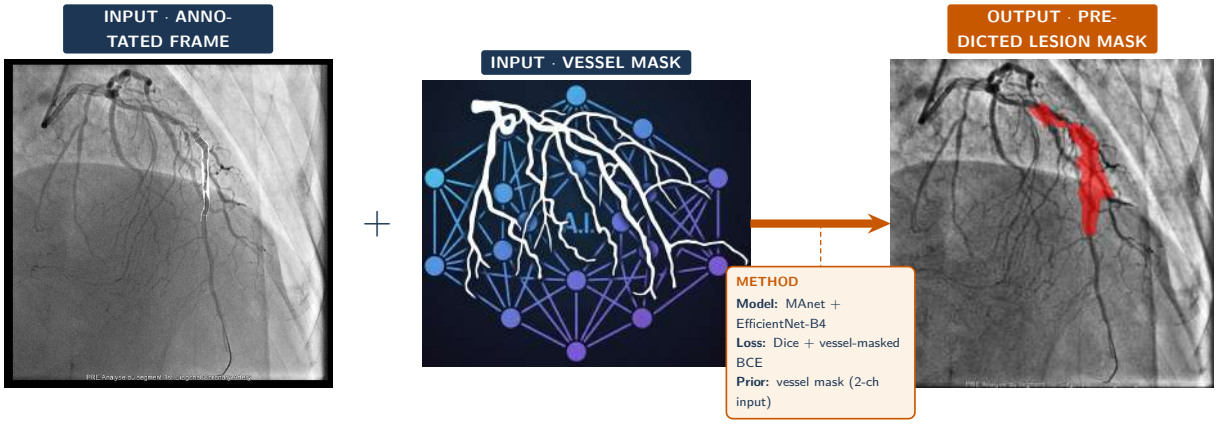


Figure 3.11: Stenosis localisation Attempt 2. The annotated frame is combined with the Phase 1 vessel-segmentation output as a two-channel input to the MAnet + EfficientNet-B4 detector. The predicted lesion mask (right) is now confined to the coronary tree — off-vessel false positives from Attempt 1 are eliminated — but the boundary remains coarse and incomplete, because the only positive supervisory signal is the weak contrast of the small annotated region.

Attempt 3 — Direct Dense Supervision (Final Model). The third attempt retained the two-channel vessel-gated input of Attempt 2 and added a manually annotated binary stenosis mask as explicit dense ground truth, providing the pixel-level boundary information that the previous attempt lacked. The annotation was produced manually in GIMP on a curated subset of the Oasis Clinic patients by clinical staff: each lesion boundary was drawn as a white (255) region on a black (0) background at the exact pixel extent of the stenotic segment.

Architecture. The final model is a **StenosisMAnet**: a MAnet decoder with an EfficientNet-B4 encoder (25,627,021 $\approx$ 25.6 M parameters), built using the `segmentation_models_pytorch` library with `encoder_weights='imagenet'`. The two-channel input (angiogram + vessel mask) is passed through the encoder to produce a multi-scale feature hierarchy; the MAnet decoder aggregates these features with multi-scale attention and produces a single-channel logit map at the 512512 input resolution.

Loss function. The training objective is a `CombinedStenosisLoss` with two components:

- **Focal Tversky Loss** (weight 0.7): with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 4/3$. The asymmetry $\alpha > \beta$ penalises false negatives (missed lesion pixels) more heavily than false positives, appropriate for a small-object segmentation task where missing the lesion is clinically more costly than a slight over-detection. The focusing exponent $\gamma = 4/3$ down-weights easy, well-segmented examples and concentrates the gradient on hard cases.
- **Vessel-Masked BCE** (weight 0.3): a binary cross-entropy loss in which pixels outside the predicted vessel mask are zeroed out, preventing the loss from penalising the network for ignoring non-coronary regions.

Formally, writing the soft prediction as $p_i = \sigma(z_i)$, the Tversky index counts the soft true positives, false negatives and false positives,

$$T = \frac{\sum_i p_i\, y_i + \epsilon}{\sum_i p_i\, y_i + \alpha \sum_i y_i\,(1 - p_i) + \beta \sum_i (1 - y_i)\, p_i + \epsilon}, \tag{3.11}$$

where $\sum_i y_i(1 - p_i)$ are the false negatives (weighted by α) and $\sum_i (1 - y_i)\, p_i$ the false positives (weighted by β). The Focal Tversky loss raises the complement of this index to the focusing exponent,

$$\mathcal{L}_{\mathrm{FT}} = (1 - T)^{\gamma}, \tag{3.12}$$

with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 4/3$ and $\epsilon = 10^{-6}$. The two components are combined as a fixed weighted sum,

$$\mathcal{L}_{\mathrm{Stenosis}} = 0.7\, \mathcal{L}_{\mathrm{FT}} + 0.3\, \mathcal{L}_{\mathrm{vBCE}}, \tag{3.13}$$

with the vessel-masked cross-entropy $\mathcal{L}_{\mathrm{vBCE}}$ as defined in Equation (3.9).

Training protocol. Training uses 5-fold cross-validation on the 200-patient cohort, with the 20 patients carrying manual stenosis annotations held out as the test set. Each fold trains for up to 200 epochs with early stopping (patience = 35 epochs, minimum improvement $\delta = 0.001$). The optimiser is **AdamW** ($\eta = 10^{-4}$, weight decay 10^{-4}) with a **CosineAnnealingWarmRestarts** scheduler ($T_0 = 50$, $T_{\mathrm{mult}} = 2$, $\eta_{\min} = 10^{-6}$). To

stabilise the encoder's pretrained features during the first few epochs, the EfficientNet-B4 encoder parameters are *frozen* for the first 10 epochs; at epoch 11 they are unfrozen and a discriminative learning rate is applied (encoder at 0.1 the decoder rate). Gradient clipping is applied at $\|\nabla\| \leq 1.0$ throughout. The batch size is 4, images are resized to 512512, and both the angiogram and the vessel mask are normalised to $[-1, 1]$ (mean 0.5, std 0.5). The data-augmentation pipeline applies, during training only: horizontal flip ($p = 0.5$), vertical flip ($p = 0.5$), random 90° rotation ($p = 0.5$), affine transformation (translation ±10%, scale $[0.8, 1.2]$, rotation ±30°; $p = 0.5$), elastic deformation ($\alpha = 120$, $\sigma = 6$; $p = 0.3$), grid distortion (5 steps, limit 0.3; $p = 0.3$), random brightness/contrast (± 0.2; $p = 0.3$), and Gaussian noise (std range $[0.02, 0.10]$; $p = 0.3$). Inference uses a CLAHE-enhanced input (clip limit 2.0, tile grid 88) with no augmentation.

Results. The five per-fold best validation Dice scores are 0.560, 0.434, 0.407, 0.501 and 0.617, giving a mean cross-validation Dice of 0.504. On the 20-patient held-out test set, evaluated at the best operating threshold of $\tau = 0.15$ identified by a threshold sweep, the model attains a mean Dice of $\mathbf{0.345 \pm 0.142}$ (median 0.354), an IoU of $\mathbf{0.217}$, a precision of $\mathbf{0.425}$ and a recall of $\mathbf{0.349}$ (Table 3.4). The low operating threshold of 0.15 reflects the fact that the network predicts low-confidence probabilities for small lesions, so a higher threshold would suppress true positives; using 0.15 maximises the Dice on the held-out set.

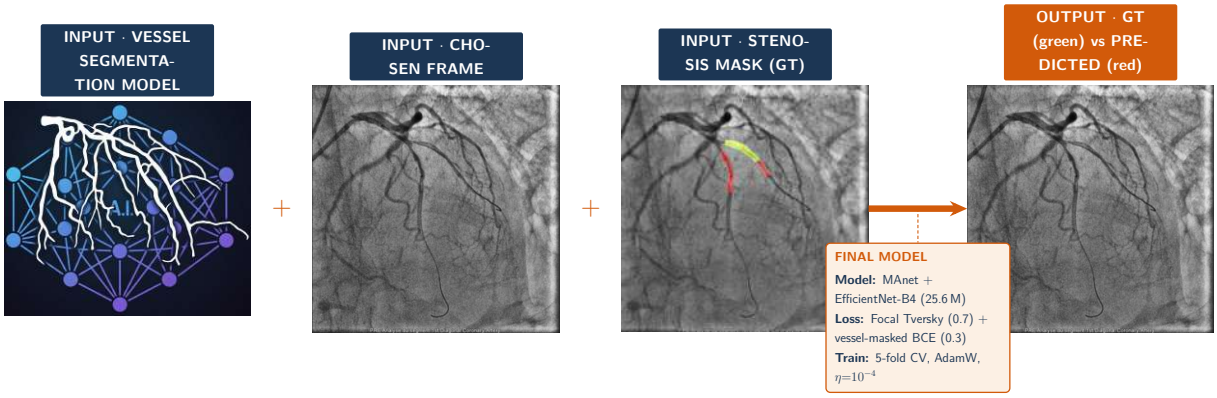


Figure 3.12: Stenosis localisation Attempt 3 (final model). Three inputs are combined: the Phase 1 vessel-segmentation model output (anatomical prior), the chosen frame, and the manually annotated stenosis ground-truth mask. The MAnet + EfficientNet-B4 detector produces the output (right): the predicted stenosis mask (red) overlaid against the ground-truth lesion boundary (green), confirming close spatial agreement with boundaries crisp enough for the downstream QCA measurement.

Phase 3 — QCA Quantification

The third stage derives the clinical QCA measurements that quantify the lesion localised by the previous stage. Its ground-truth values are read directly from the private DICOM

tags written by the angiography console (Section 3.3.6): tag `(0009,1010)` carries the vessel identifier (LAD, LCx, RCA or LM), `(0009,1013)` the reference vessel diameter (RVD), `(0009,1014)` the minimal lumen diameter (MLD), and `(0009,1016)` the lesion length. The percentage diameter stenosis is obtained from these as $\%DS = (1 - \mathrm{MLD}/\mathrm{RVD})100$ and is interpreted on the usual clinical severity scale — below 50% non-significant, 50–70% intermediate, and above 70% significant.

The quantification model itself extends the stenosis network with a second output: it is a MAnet with an EfficientNet-B4 encoder that retains the Focal Tversky plus vessel-masked BCE segmentation objective and adds a mean-squared-error regression head for the two QCA targets, so that the lesion mask and its numerical descriptors are produced together in one model. The regression head is trained with a mean-squared-error objective over the two QCA targets,

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{r}_k - r_k \right)^2, \tag{3.14}$$

where r_k and $\hat{r}_k$ are the ground-truth and predicted values of the $K = 2$ targets (the percentage diameter stenosis $\%DS$ and the lesion length). The full objective of this stage augments the segmentation loss of Equation (3.13) with this regression term under a pair of annealed task weights,

$$\mathcal{L}_{\mathrm{QCA}} = w_{\mathrm{seg}} \left(0.7\, \mathcal{L}_{\mathrm{FT}} + 0.3\, \mathcal{L}_{\mathrm{vBCE}} \right) + w_{\mathrm{reg}}\, \mathcal{L}_{\mathrm{MSE}}, \tag{3.15}$$

where the weight pair $(w_{\mathrm{seg}}, w_{\mathrm{reg}})$ is the concrete instantiation of the balancing constant referred to above: rather than a fixed trade-off, the regression weight is ramped up in three stages,

$$(w_{\mathrm{seg}}, w_{\mathrm{reg}}) = \begin{cases} (0.95,\ 0.05) & \text{epochs 1–50,} \\ (0.85,\ 0.15) & \text{epochs 50–150,} \\ (0.80,\ 0.20) & \text{epochs } > 150, \end{cases}$$

so the shared encoder is not pulled toward the QCA targets before the lesion mask is reliable. Importantly, this combined segmentation-and-regression behaviour is internal to the stage; it does not couple back into the vessel or stenosis stages, which remain separately trained, so the pipeline as a whole stays decoupled.

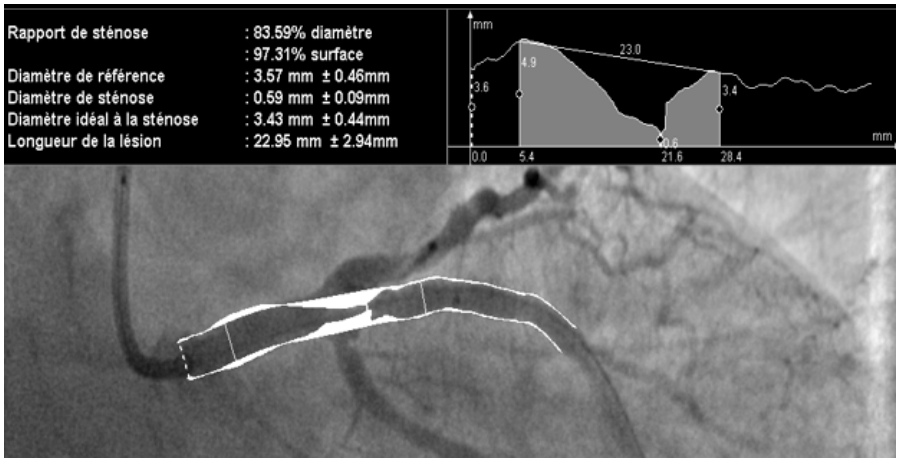


Figure 3.13: Example QCA quantification output for the third stage. The diseased segment is delineated along the vessel and the console reports the derived clinical measurements — reference vessel diameter (RVD), minimal lumen diameter (MLD), the resulting percentage diameter stenosis ($\%DS$), and the lesion length — which serve as the ground-truth targets read from the private DICOM tags. In the case shown the lesion reaches a $\%DS$ in the diagnostically significant range ($> 70\%$).

Flow 1 — Quantitative Results

The three stages are evaluated independently, since no gradient couples them. The vessel stage carries no withheld manual masks at this stage of development and is therefore reported by its converged training loss (0.045); a quantitative held-out vessel Dice is reported later, in the cross-flow comparison of Section 3.12. The stenosis and QCA stages are both evaluated on the same 20-patient held-out set. The standalone stenosis detector reaches a mean Dice of 0.345 ± 0.142 (median 0.354), with an IoU of 0.217, a precision of 0.425 and a recall of 0.349. The QCA stage, whose segmentation head is trained jointly with the regression head, reaches a segmentation Dice of 0.465 ± 0.326 (median 0.563), a $\%DS$ MAE of 18.9%, and a lesion-length MAE of 5.7 mm. The complete results are summarised in Table 3.4, and the per-epoch training dynamics of all three stages are shown in Figure 3.14.

Table 3.4: Quantitative results of the Flow 1 sequential pipeline. The vessel stage carries no withheld manual masks and is reported by its converged training loss; the stenosis and QCA stages are evaluated on $n = 20$ held-out patients at the tuned operating threshold $\tau = 0.15$. The stenosis Dice differs between Phase 2 (0.345) and the Phase 3 segmentation head (0.465) because they are two separately trained models.

Phase	Model	Metric	Value	n
1 — Vessel	U-Net (CycleGAN)	Train loss	0.045	—
2 — Stenosis	MAnet + EfficientNet-B4	Mean Dice	0.345 ± 0.142	20
		Median Dice	0.354	20
		IoU	0.217	20
		Precision	0.425	20
		Recall	0.349	20
3 — QCA	MAnet + EfficientNet-B4	Seg. Dice (mean)	0.465 ± 0.326	20
		Seg. Dice (median)	0.563	20
		$\%DS$ MAE	18.9%	20
		Lesion-length MAE	5.7 mm	20

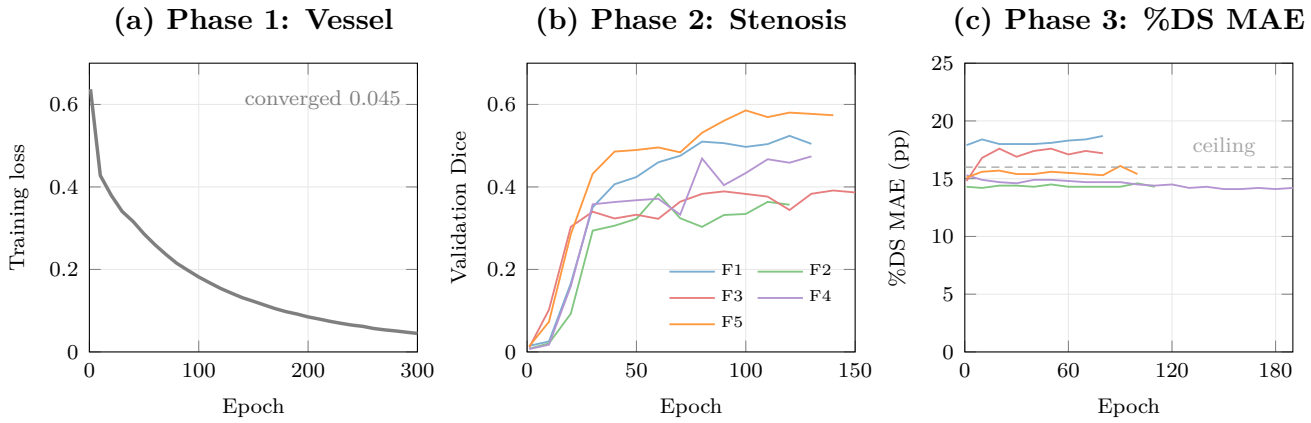


Figure 3.14: Training dynamics of the Flow 1 sequential pipeline, plotted from the per-epoch logs of the three independently trained notebooks (each colour is one of the five cross-validation folds, F1–F5). **(a)** Phase 1 vessel U-Net training loss over 300 epochs, decaying smoothly to 0.045. **(b)** Phase 2 stenosis validation Dice (logging threshold 0.30); the wide fold spread (0.36–0.62 at best) reflects the high variance later summarised as the held-out 0.345 ± 0.142 at the tuned threshold 0.15 (Table 3.4). **(c)** Phase 3 QCA $\%DS$ mean absolute error: although the segmentation Dice keeps rising during these same runs, the regression error stays pinned in the 14–18 pp band across every fold and every epoch. This flat error against improving segmentation is the visual signature of the cascading-error ceiling — the upstream mask can no longer help the downstream measurement — which is precisely the limitation that motivates the jointly optimised Flow 2.

Flow 1 — The Cascading-Error Limitation

The defining weakness of the sequential pipeline is structural rather than incidental. Because each stage is trained against its own objective and no gradient flows backward across stage boundaries, an error introduced in Phase 1 is amplified and passed uncorrected to Phase 2, which in turn passes a compounded error to Phase 3. The consequence is visible in the training dynamics of Figure 3.14(c): the $\%DS$ regression error remains pinned in the 14–18 pp band regardless of how much the upstream segmentation improves, because no amount of downstream optimisation can recover information that an earlier stage has already discarded. With a held-out Phase 2 Dice of only 0.345, the residual mask error sets a hard ceiling on the QCA accuracy that Phase 3 can attain. This cascading-error ceiling is precisely the limitation that motivates the jointly optimised single-network design of Flow 2, in which a shared encoder allows the three tasks to regularise one another and the QCA head to read the lesion location directly.

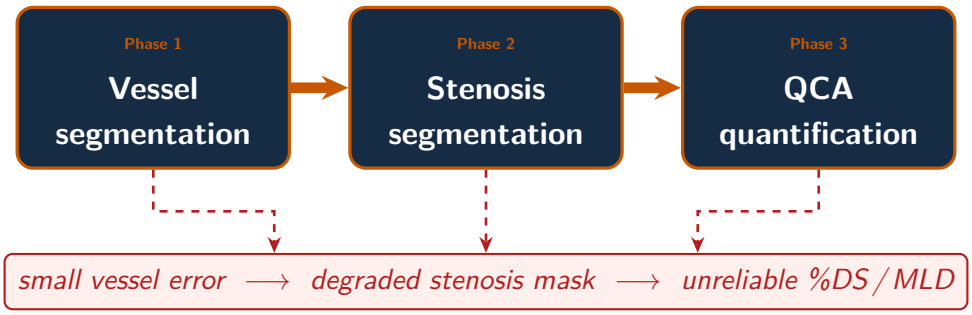


Figure 3.15: The cascading-error limitation of the Flow 1 sequential pipeline. Each stage is trained independently, so an error introduced in Phase 1 (vessel segmentation) is amplified and passed uncorrected to Phase 2 (stenosis segmentation), which passes a compounded error to Phase 3 (QCA quantification). The residual mask error sets a hard ceiling on Phase 3 that no downstream optimisation can recover.

3.6 Flow 2 — Single-Source / Multi-Task Architecture: GCT-Net

3.6.1 Design Rationale

GCT-Net is organised around four architectural choices that together implement the single-source multi-task formulation of Flow 2:

1. *Hybrid CNN-Transformer encoder.* A pure CNN encoder captures local vessel texture but cannot model the topology of the coronary tree, whereas a pure Vision Transformer requires too much data to learn convolutional inductive biases from scratch in a single-centre setting. MaxViT-tiny [15] (Section 3.6.4) provides a middle ground: each MaxViT block alternates an MBConv layer (local), a block atten-

tion (windowed local) and a grid attention (sparse global), yielding both locality and a global receptive field with linear complexity. The same backbone family has previously been adopted in coronary CT analysis [27], supporting its relevance to vascular imaging.

2. *Hard parameter sharing across the three tasks.* A single shared encoder feeds three task-specific heads. With only $\sim$ 80 patients carrying a manual stenosis mask, three independent encoders would overfit; sharing forces a common representation to serve all three tasks and acts as an implicit regulariser (Section 1.6.4) [17, 18, 23].

3. *Decoupling the dense prediction head from the regression head with a stop-gradient.* The QCA head is conditioned on the predicted stenosis map, but the gradient of the QCA loss is blocked at this junction by a stop-gradient operation (`.detach()`), so that the QCA objective can read the spatial content of the predicted lesion without being able to reshape the localisation map (Section 3.6.7).

4. *Lesion-aware ROI pooling for the QCA head.* Rather than globally pooling the bottleneck — which dilutes the small stenotic signal — the QCA head pools encoder features inside the predicted stenosis region, using the detached stenosis map as a soft spatial attention mask (Section 3.6.6). This is the single most important architectural fix for accurate $\%DS$ and lesion-length estimation.

The chosen ICA frame consumed by the encoder is itself produced automatically by a best-frame selection module that re-uses the Phase 1 vessel model as a teacher (Section 3.6.3); no manual frame-selection step is required at deployment. A schematic of the architecture is shown in Figure 3.16.

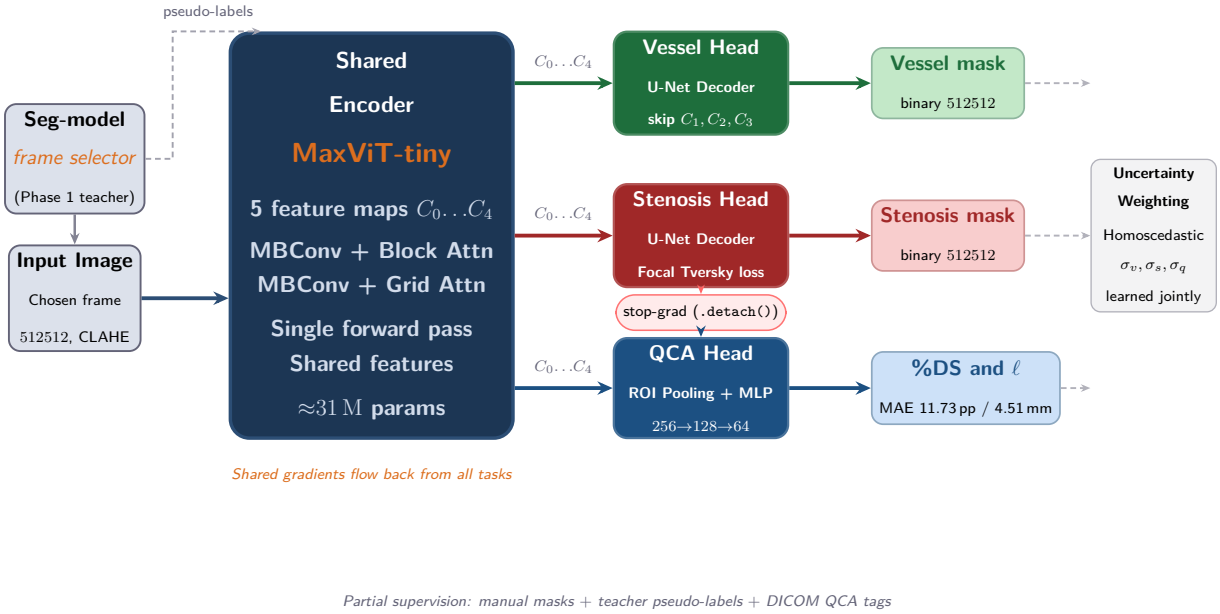


Figure 3.16: Overview of GCT-Net (Flow 2). A shared MaxViT-tiny encoder receives the chosen ICA frame and produces five hierarchical feature maps in a single forward pass. Three task-specific heads share these features: a U-Net vessel decoder, a U-Net stenosis decoder, and a lesion-aware ROI pooling QCA head. A hard stop-gradient (`.detach()`) between the stenosis decoder and the QCA head prevents regression gradients from corrupting the localisation map. Task losses are balanced by homoscedastic uncertainty weighting $(\sigma_v, \sigma_s, \sigma_q)$.

3.6.2 Novel Contributions

Three of the design choices above are not merely engineering decisions but constitute the core methodological contributions of this work. Each one addresses a concrete failure of the naive single-source multi-task baseline, and each is presented here following the same problem–mechanism–consequence structure: the problem it solves, how the mechanism works (with its governing equations), and why it matters for the final system.

Contribution 1 — Lesion-Aware ROI Pooling

The problem. A stenotic lesion occupies less than 1% of the angiographic frame. A direct regression head that global-average-pools the encoder bottleneck therefore averages the small diagnostic signal together with the entire background, so the lesion is numerically drowned out and the resulting $\%DS$ and lesion-length estimates are unreliable. Pooling must be focused on the lesion, not the frame.

How it works. Instead of pooling globally, the QCA head pools the encoder features *only inside the predicted lesion*, using the detached stenosis probability map m as a soft spatial-attention weight. For each encoder scale i, the attention-weighted (soft-pooled)

feature is

$$f_i^{\text{pool}} = \frac{\sum_{u,v} C_i[u,v] \cdot m_i[u,v]}{\sum_{u,v} m_i[u,v] + \varepsilon}, \tag{3.16}$$

where $C_i[u,v]$ is the encoder feature at spatial location (u,v) of scale i, m_i is the stenosis mask resized to that scale, and ε is a small constant preventing division by zero. The numerator accumulates feature energy only where the mask is active and the denominator normalises by the mask's total mass, so Equation (3.16) computes the average feature *within the lesion region* rather than over the whole frame.

A difficulty arises when the predicted mask is nearly empty: the denominator $\sum_{u,v} m_i[u,v]$ collapses toward zero and the soft-pooled feature becomes numerically unstable. This is resolved with an empty-mask fallback that blends the soft pool with a conventional global-average pool $g_i = \text{GAP}(C_i)$ according to the mask's peak confidence $\rho_i = \max_{u,v} m_i[u,v] \in [0,1]$:

$$\tilde{f}_i^{\text{pool}} = \rho_i \, f_i^{\text{pool}} + (1-\rho_i)\, g_i. \tag{3.17}$$

The interpolation in Equation (3.17) makes the behaviour adaptive to the network's own confidence. When the network is confident in the lesion location, $\rho_i \to 1$ and the expression reduces to the lesion-only soft pool; when it predicts no lesion, $\rho_i \to 0$ and the expression reverts smoothly to a global average, so the head always receives a well-defined feature vector. In addition to the pooled features, a five-dimensional vector of differentiable geometric statistics — lesion area fraction, centroid coordinates (c_y, c_x), and principal/secondary spatial axes (p,q) — is extracted from the same mask, giving the regression head explicit access to the size, position, and elongation of the lesion.

Why it matters. Measuring features at the lesion itself lets the QCA head observe the stenosis directly rather than through a diluted frame-wide average. This is the single most important architectural change for accurate $\%DS$ and lesion-length estimation: it is the mechanism that allows a regression head, attached to a segmentation backbone, to recover clinically meaningful continuous measurements.

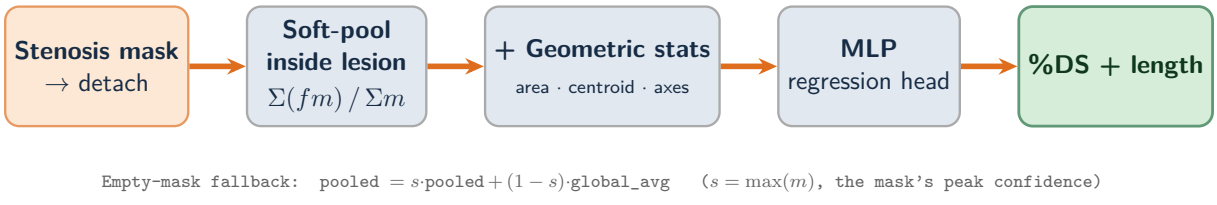


Figure 3.17: Contribution 1 — lesion-aware ROI pooling pipeline. The detached stenosis mask drives a soft pool that averages encoder features inside the lesion only ($\Sigma(fm)/\Sigma m$); geometric statistics (area, centroid, axes) are appended, and an MLP regresses $\%DS$ and lesion length. The empty-mask fallback blends the soft pool with a global average according to the mask's peak confidence $s = \max(m)$.

Contribution 2 — Stop-Gradient: Protecting the Stenosis Map

The problem. Conditioning the QCA head on the predicted stenosis map is exactly what gives the regression its focused signal — but it opens a dangerous back-channel. If the QCA loss is allowed to back-propagate through the stenosis prediction, the optimiser can take a shortcut: rather than improving the regression, it can *warp the stenosis mask* into whatever shape makes $\%DS$ and length easiest to predict, regardless of whether that shape corresponds to the true lesion. The localisation map would then be silently corrupted by the regression objective.

How it works. The coupling is made strictly one-directional with a hard stop-gradient. The soft mask consumed by the lesion-aware ROI pool is

$$m \;=\; \sigma\big(\text{stop\_grad}(s)\big), \tag{3.18}$$

where $s \in \mathbb{R}^{B1HW}$ are the stenosis logits (B the mini-batch size, HW the spatial grid), $\sigma(\cdot)$ is the sigmoid, and stop_grad$(\cdot)$ is the `.detach()` operation. The forward pass is unchanged — the QCA head still receives the complete spatial content of the predicted lesion. The backward pass, however, carries *no* QCA gradient into the stenosis decoder: in Equation (3.18) the detach severs the gradient path at the mask, so $\partial \mathcal{L}_{\text{qca}}/\partial s = 0$. The stenosis map is therefore shaped only by its own segmentation loss and by the shared encoder.

Why it matters. The QCA head is spatially conditioned on the lesion location, yet can never corrupt it. The stenosis map stays clean and anatomically faithful while still driving the regression. (Flow 3 later relaxes this hard cut into an annealable coupling that permits a small, controlled gradient leakage in its final stage; see Section 3.11.)

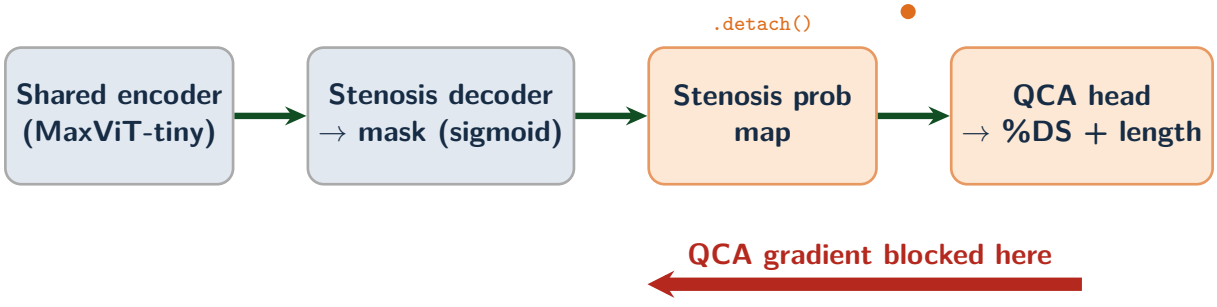


Figure 3.18: Contribution 2 — the stop-gradient that protects the stenosis map. The forward path (green) runs shared encoder → stenosis decoder → stenosis probability map → QCA head, so the QCA head still reads the full lesion content. The `.detach()` at the probability-map→QCA junction blocks the backward path (red): the QCA gradient cannot flow back into the stenosis decoder, so the regression can never reshape the localisation map.

Contribution 3 — Shared Encoder and Self-Balancing Loss

The problem. Two coupled difficulties arise in a single-centre, small-data multi-task setting. First, training three independent encoders — one per task — on roughly 80 labelled patients would overfit badly. Second, the three task losses have different magnitudes and noise levels, and hand-tuning fixed weights ($\lambda_{\text{ves}}, \lambda_{\text{sten}}, \lambda_{\text{qca}}$) to balance them requires a costly grid search that does not transfer across datasets.

How it works. A single shared MaxViT-tiny encoder feeds all three task heads, so the bulk of the parameters are trained on every patient regardless of which labels that patient carries; the shared representation acts as an implicit regulariser. The loss weights are then made *self-balancing* through homoscedastic uncertainty weighting [16]: each task t is assigned a learnable log-variance $\log \sigma_t^2$, optimised jointly with the network. Writing the precision $\tau_t = \exp(-\log \sigma_t^2) = 1/\sigma_t^2$, the total objective is

$$\mathcal{L}_{\text{total}} \;=\; \sum_t \Big(\tau_t \, \mathcal{L}_t + \log \sigma_t^2 \Big), \tag{3.19}$$

summed over the vessel, stenosis, and QCA tasks. In Equation (3.19), the precision τ_t scales each task loss, so a task the network finds uncertain (large σ_t) is automatically down-weighted, while the regulariser $\log \sigma_t^2$ prevents the trivial solution of driving all variances to zero. The weights are not hand-set: the network learns how much to trust each task directly from the data.

Why it matters. A single encoder regularises the small dataset and removes the parameter blow-up of per-task encoders, while the learned uncertainties automatically down-weight the noisier QCA task (whose targets are parsed from DICOM tags) relative to the cleaner pixel masks — with no grid search over loss weights. The detailed loss formulation, including the partial-supervision validity flags and the precise form of each task loss, is developed in Section 3.7.

3.6.3 Automatic Best-Frame Selection

Each patient in the Oasis Clinic dataset is represented by several DICOM files under the `DICOM/PA<id>/ST0/input/` path, each storing a multi-frame cine sequence of 15–183 frames. A frame is clinically usable for quantitative analysis only during a narrow temporal window of the cine: at the peak of the iodinated-contrast injection, when the vessel tree is fully opacified, and away from the systolic phase, where rapid cardiac motion produces blur. The remaining frames are dominated by partially opacified vessels, motion artefacts, or the pre- and post-injection phases in which contrast is absent. The original Oasis acquisition workflow side-stepped this by relying on a single frame manually selected

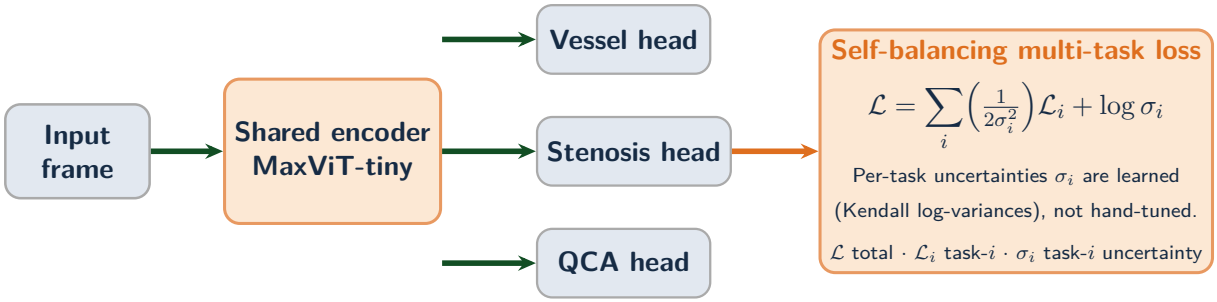


Figure 3.19: Contribution 3 — a single shared encoder with a self-balancing multi-task loss. The input frame passes through one shared MaxViT-tiny encoder that feeds all three task heads (vessel, stenosis, QCA). The three task losses are combined by homoscedastic uncertainty weighting, in which the per-task uncertainties σ_i are learned (Kendall log-variances) rather than hand-tuned, so the noisy QCA task is automatically down-weighted without a grid search.

by clinical staff (the `chosen_image` convention), which limits the usable data to one image per patient and introduces an additional source of human variability.

To remove this manual bottleneck, an automatic best-frame selection module is integrated into the preprocessing pipeline; it re-uses the Phase 1 vessel segmentation model as a teacher to score candidate frames. The module is implemented as a five-stage pipeline (Figure 3.20): the multi-frame DICOM files are decoded, the teacher scores every sampled frame by its vesselness response, per-frame vesselness maps are produced, a composite score combining vessel area, intensity, contrast and a blur penalty is computed, and the frames are ranked. The rank-1 frame is forwarded to the multi-task network, while ranks 2–3 are cached as a deterministic fallback.

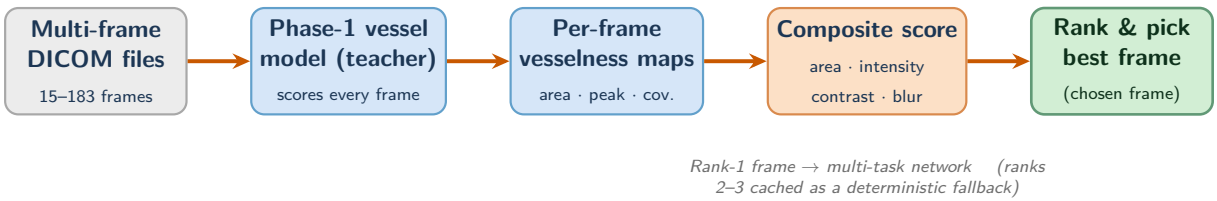


Figure 3.20: The automatic best-frame selection pipeline. Multi-frame DICOM files are decoded; the Phase 1 vessel segmentation model, re-used as a teacher, scores every sampled frame by its vesselness response; per-frame vesselness maps yield a composite score combining vessel area, intensity, contrast and a blur penalty; and the frames are ranked so that the rank-1 frame is forwarded to GCT-Net, with ranks 2–3 cached as a deterministic fallback.

Candidate sampling. Because a long cine may contain up to 183 frames and scoring every one is wasteful, the selector first samples at most 30 frames by uniform temporal spacing (`numpy.linspace` over the frame indices, duplicates removed), so that early, middle, and late portions of the injection are all represented. Each candidate is then preprocessed identically before scoring: greyscale conversion, min–max normalisation to

$[0, 1]$, conversion to 8-bit, CLAHE contrast enhancement (clip limit 2.0, tile grid 88), a 33 Gaussian blur to suppress noise that would mislead the sharpness measure, and a resize to 512512 to match the model input.

Frame-level scoring cues. For a CLAHE-enhanced frame $I \in [0, 255]^{HW}$, five cues are computed and combined: two model-independent image-quality cues and three task-relevant cues derived from the teacher U-Net f_θ. The image-quality cues are the *Laplacian-variance sharpness* $S_{\text{sharp}}(I) = \text{Var}(\nabla^2 I)$, which collapses under the motion blur of the systolic phase and so favours sharp frames, and the *contrast spread* $S_{\text{contrast}}(I) = \text{std}(I)/255$, which is near-zero for pre-injection frames and large for fully-opacified ones. The three teacher-driven cues query f_θ at scoring time:

$$S_{\text{area}}(I) = \frac{1}{HW} \sum_{u,v} \sigma\big(f_\theta(I)\big)[u, v], \qquad S_{\text{peak}}(I) = \max_{u,v} \sigma\big(f_\theta(I)\big)[u, v], \tag{3.20}$$

$$S_{\text{cov}}(I) = \frac{1}{HW} \sum_{u,v} \mathbb{1}\big[\sigma\big(f_\theta(I)\big)[u, v] > 0.5\big], \tag{3.21}$$

where σ is the sigmoid, $f_\theta(I)$ is the teacher logit map, (u, v) index the spatial grid, and $\mathbb{1}[\cdot]$ is the indicator function. The vessel-area cue S_{area} favours frames in which the whole tree is opacified, the peak cue S_{peak} favours strong local contrast filling, and the coverage cue S_{cov} favours frames in which a continuous vascular network is recoverable at threshold 0.5.

Score aggregation. Acquisition exposure varies between studies, so absolute thresholds on any single cue are not robust. All cues are instead normalised to $[0, 1]$ within each patient by min–max scaling over that patient's candidate frames, and linearly combined:

$$S(I) = 0.40\, \tilde{S}_{\text{area}} + 0.20\, \tilde{S}_{\text{peak}} + 0.20\, \tilde{S}_{\text{cov}} + 0.10\, \tilde{S}_{\text{contrast}} + 0.10\, \tilde{S}_{\text{sharp}}, \tag{3.22}$$

where $\tilde{\cdot}$ denotes the patient-normalised cue. The weights in Equation (3.22) prioritise direct task-relevance (the teacher vessel quantities, 0.80 of the total weight) over generic image quality, reflecting the fact that a slightly blurry but fully-opacified frame is more useful than a perfectly sharp but empty one. For each patient the top-scoring frames are stored in a JSON cache that maps the patient identifier to a ranked list of records, built once per dataset and consulted by the dataset class.

Use at training and inference. The cache supports two access modes. In the training and validation mode reported in this chapter, the manually selected `chosen_image` is used for each patient, because the manual ground-truth masks were drawn on that specific frame and would not align with a frame selected elsewhere in the cine; the auto-

frame cache is consulted only as a diagnostic. In the deployment mode used by the web application (Section 3.14), `chosen_image` is discarded entirely and the rank-1 auto-scored frame drives the pipeline, with ranks 2–3 available as a fallback if the rank-1 frame fails a downstream sanity check. The deployment pipeline is therefore no longer dependent on the manual frame-selection step: the system consumes raw DICOM archives without operator intervention, in line with the goal of an end-to-end automated system.

3.6.4 Shared MaxViT Encoder

The shared encoder is the `maxvit_tiny_tf_512.in1k` variant of MaxViT [15] ($\sim 31\,\mathrm{M}$ parameters), pretrained on ImageNet at 512512 resolution and loaded through the `timm` library [19] in `features_only` mode. It is the single most important component of GCT-Net: it is the only part of the network shared across all three tasks, and the five feature maps it produces are the sole point of contact between the vessel, stenosis, and QCA heads.

Building Blocks of a MaxViT-tiny Stage

The motivation for choosing MaxViT over a pure CNN or a pure Vision Transformer is that coronary analysis needs *both* fine local texture (to trace thin, low-contrast vessels) and a global receptive field (to reason about the topology of the whole coronary tree). MaxViT obtains both within a single stage by chaining one convolutional block with two complementary attention mechanisms:

- **MBConv (local).** An inverted-residual mobile convolution block, the same building block used in EfficientNet [3]. It applies a depthwise convolution with a squeeze-and-excitation gate, giving the stage a strong convolutional inductive bias that captures fine vessel textures with very few parameters. This is the component that lets the encoder learn local edge and ridge structure even on a small dataset.
- **Block attention (windowed local).** Self-attention computed *within* non-overlapping P P windows. Each window attends only to itself, so the cost is linear in the number of pixels while still modelling patch-level interactions that a convolution of the same receptive field cannot. This captures medium-range dependencies between nearby vessel segments.
- **Grid attention (sparse global).** Self-attention computed along a sparse, dilated grid that spans the entire feature map. By attending to a regularly spaced subset of positions rather than all pairs, grid attention achieves a *global* receptive field while keeping the cost linear in the image size — the key property that makes a transformer affordable at 512512 resolution. This is what lets the encoder relate distant branches of the coronary tree to one another.

Within each MaxViT-tiny stage these three operations are applied in sequence, so that a tensor passes through local convolution, then windowed attention, then sparse-global attention before leaving the stage. The chaining is illustrated in Figure 3.21.

Figure 3.21: How the three blocks chain inside a single MaxViT-tiny stage. A tensor passes through an MBConv (local inverted-residual convolution), then block attention (self-attention within non-overlapping PP windows), then grid attention (sparse self-attention spanning the full feature map), combining a local and a global receptive field within one block at linear cost.

Hierarchical Feature-Map Output

Given a 3512512 input, the encoder produces a list of five hierarchical feature maps $[C_0, C_1, C_2, C_3, C_4]$ at decreasing spatial resolutions, with strides $1/2, 1/4, 1/8, 1/16, 1/32$ relative to the input. The shallow maps (C_0, C_1) retain high spatial resolution and encode fine local detail such as vessel edges, while the deep maps (C_3, C_4) are spatially coarse but semantically rich, encoding the global layout of the coronary tree. This multi-scale pyramid is exactly what the U-Net decoders consume through their skip connections, and it is what the lesion-aware QCA head pools over.

Crucially, these five tensors are the *only* point of contact between the three task heads: the vessel head, the stenosis head, and the QCA head each receive the same $[C_0, \ldots, C_4]$ and never exchange information except through this shared representation. The same MaxViT backbone family has previously been applied to coronary CT analysis by Gerbasi et al. [27], supporting its relevance to vascular imaging. The hierarchy and its sharing are illustrated in Figure 3.22.

3.6.5 Vessel and Stenosis Decoders

The vessel and stenosis tasks are both pixel-wise binary classification problems, so the same decoder topology is used for each: a five-level U-Net decoder [5] with skip connections from each encoder scale. Each decoder block performs (i) bilinear 2 up-sampling, (ii) concatenation with the encoder skip at that resolution, and (iii) a double convolution composed of two 33 convolutions, each followed by Group Normalisation (with eight groups) and a SiLU activation. Decoder channel widths follow the schedule $(256, 128, 64, 32, 16)$, ending with a 11 convolution producing a single-channel logit map at the input resolution; if the deepest feature map is downsampled by an additional factor of two relative to the lowest decoder level, the logits are bilinearly interpolated to the input resolution. The

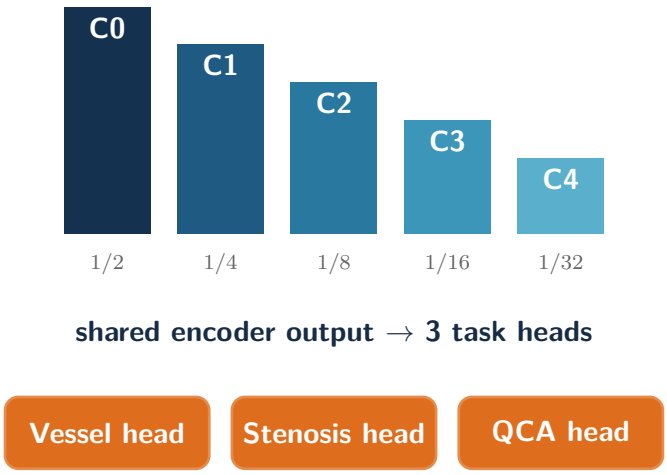


Figure 3.22: The five hierarchical feature maps emitted by the shared MaxViT-tiny encoder. From C_0 (stride 1/2, high spatial resolution, fine local detail) to C_4 (stride 1/32, spatially coarse, semantically global), the pyramid spans both ends of the scale axis. These five maps are shared, unchanged, as the sole input to all three task heads — vessel, stenosis, and QCA.

two decoders do not share parameters with each other; they share gradients only through the encoder.

3.6.6 Lesion-Aware QCA Head

The QCA head differentiates the proposed model from the segmentation networks reviewed in Chapter 2 by producing continuous lesion measurements directly, rather than as a post-processing step. A direct approach to QCA regression would global-average-pool the deepest encoder feature map and feed the result to an MLP, but this produces unstable estimates because a small stenosis occupies a tiny fraction of the image and its signal is diluted by the average. The mechanism adopted here is a *lesion-aware ROI pooling* that uses the predicted stenosis map itself as a spatial attention mask over the encoder features. Conceptually, this extends two established ideas: hard region-of-interest pooling on predicted bounding boxes, used in Mask R-CNN [6] for instance segmentation, and mask-conditioned feature gating, used in Attention U-Net and related medical-imaging architectures [54], to a continuous regression target driven by a soft segmentation map. Given the stenosis logits $s \in \mathbb{R}^{B1512512}$, where B is the mini-batch size and the remaining dimensions are the single logit channel and the 512512 spatial grid, the head proceeds as follows:

1. *Detach and resize.* The soft mask $m = \sigma(\text{stop\_grad}(s))$ is computed (the stop-gradient is critical, see Section 3.6.7), then bilinearly resized to the spatial size of each encoder scale C_i.

2. *Multi-scale soft pooling with empty-mask fallback.* For each encoder scale i, an

attention-weighted feature is computed as

$$f_i^{\text{pool}} \;=\; \frac{\sum_{u,v} C_i[u,v] \cdot m_i[u,v]}{\sum_{u,v} m_i[u,v] + \varepsilon}. \tag{3.23}$$

When the predicted stenosis mask is nearly empty, the denominator becomes very small and the soft-pooled feature is unstable. To avoid this, the soft pool is blended with a global-average pool $g_i = \text{GAP}(C_i)$ according to the maximum mask intensity $\rho_i = \max_{u,v} m_i[u,v] \in [0,1]$:

$$\tilde{f}_i^{\text{pool}} \;=\; \rho_i \, f_i^{\text{pool}} + (1-\rho_i)\, g_i. \tag{3.24}$$

When the network is confident in the lesion location, $\rho_i \to 1$ and the soft pool dominates; when it predicts no lesion, $\rho_i \to 0$ and the pool reverts smoothly to a global average.

3. *Differentiable geometric statistics.* From the same soft mask m, a five-dimensional vector $g = (a, c_y, c_x, p, q)$ is computed, where a is the area fraction of m, (c_y, c_x) are the normalised centroid coordinates, and p, q are the principal and secondary spatial standard deviations, defined as $p = \sqrt{\max(\text{var}_y, \text{var}_x) + \varepsilon}$ and $q = \sqrt{\min(\text{var}_y, \text{var}_x) + \varepsilon}$. These statistics give the head explicit access to the size, position and elongation of the predicted lesion.

4. *Projection and regression.* The pooled features $\left[\tilde{f}_0^{\text{pool}}, \ldots, \tilde{f}_4^{\text{pool}}, g\right]$ are concatenated and passed through a linear projection (Linear – LayerNorm – SiLU) to a 256-dimensional bottleneck. A small MLP with two hidden layers ($256 \to 128 \to 64$) and dropout regularisation ($p = 0.3$ then 0.2) outputs two values which are passed through a sigmoid to lie in $[0,1]^2$, namely the predicted stenosis percentage divided by 100 and the predicted lesion length divided by 50 mm.

This design ensures that the regression depends spatially on what the stenosis decoder predicts, without coupling their gradients (next section).

3.6.7 Stop-Gradient between the Stenosis Decoder and the QCA Head

Conditioning the QCA head on the predicted stenosis map raises a tension between two objectives. On the one hand, conditioning the regression on the lesion location is exactly what gives the QCA head a focused, informative signal. On the other hand, if the QCA gradient is allowed to flow back through the stenosis prediction, the QCA loss can pull the stenosis map toward a solution that minimises the regression error at the expense of correct localisation — a degenerate shortcut in which the network distorts the lesion mask

simply to make $\%DS$ and length easier to predict, regardless of whether the highlighted region corresponds to the actual lesion.

Flow 2 resolves this with a hard stop-gradient. The soft mask consumed by the lesion-aware ROI pool is

$$m \;=\; \sigma\big(\text{stop\_grad}(s)\big), \tag{3.25}$$

where $s \in \mathbb{R}^{B1HW}$ are the stenosis logits (with B the mini-batch size, HW the spatial grid) and stop_grad$(\cdot)$ is implemented by the `.detach()` operation. The forward pass is unaffected: the QCA head still receives the full spatial content of the predicted lesion. The backward pass, however, carries no QCA gradient into the stenosis decoder, so the localisation map is shaped only by its own segmentation loss and by the shared encoder. The stenosis map therefore stays clean and accurate while still spatially conditioning the regression. (Flow 3 relaxes this hard separation into an annealable coupling that permits a small, controlled gradient leakage in its final stage; see Section 3.11.)

3.7 Multi-Task Loss with Homoscedastic Uncertainty Weighting

The total loss balances three task-specific objectives whose magnitudes, gradient scales and label-noise levels differ substantially. Hand-tuning fixed weights $(\lambda_{\text{ves}}, \lambda_{\text{sten}}, \lambda_{\text{qca}})$ is impractical; the implementation therefore uses homoscedastic uncertainty weighting [16] (Section 1.6.4). Three learnable log-variances $(\log\sigma_v^2, \log\sigma_s^2, \log\sigma_q^2)$ are optimised jointly with the network parameters. Writing the precision as $\tau_t = \exp(-\log\sigma_t^2) = 1/\sigma_t^2$, the total loss takes the form

$$\mathcal{L}_{\text{total}} \;=\; \tau_v\,\mathcal{L}_{\text{ves}} + \log\sigma_v^2 + \mathbb{1}_s\big(\tau_s\,\mathcal{L}_{\text{sten}} + \log\sigma_s^2\big) + \mathbb{1}_q\big(\tau_q\,\mathcal{L}_{\text{qca}} + \log\sigma_q^2\big), \tag{3.26}$$

where $\mathbb{1}_s$ and $\mathbb{1}_q$ are batch-level indicator variables equal to one when the batch contains at least one sample with a valid stenosis mask or QCA target, respectively. The conditional inclusion is necessary because, when a task carries no signal in a given mini-batch, including its log-variance term in the total loss would induce a non-informative gradient that would push $\log\sigma_t^2$ towards zero and inflate the effective weight of the task. The regularisation term $\log\sigma_t^2$ prevents the variances of active tasks from collapsing to zero. Tasks with noisier labels (here, QCA targets parsed from DICOM tags) automatically receive a larger σ_t and therefore a smaller effective weight, without manual tuning. The log-variances are initialised to $(0, 0, 1)$, biasing the early training towards the cleaner segmentation losses.

The three task losses are defined below. Throughout, the per-sample validity flags $w_b^v \in \{0, 0.5, 1\}$, $w_b^s \in \{0, 1\}$ and $w_b^q \in \{0, 1\}$ implement partial supervision by zeroing the

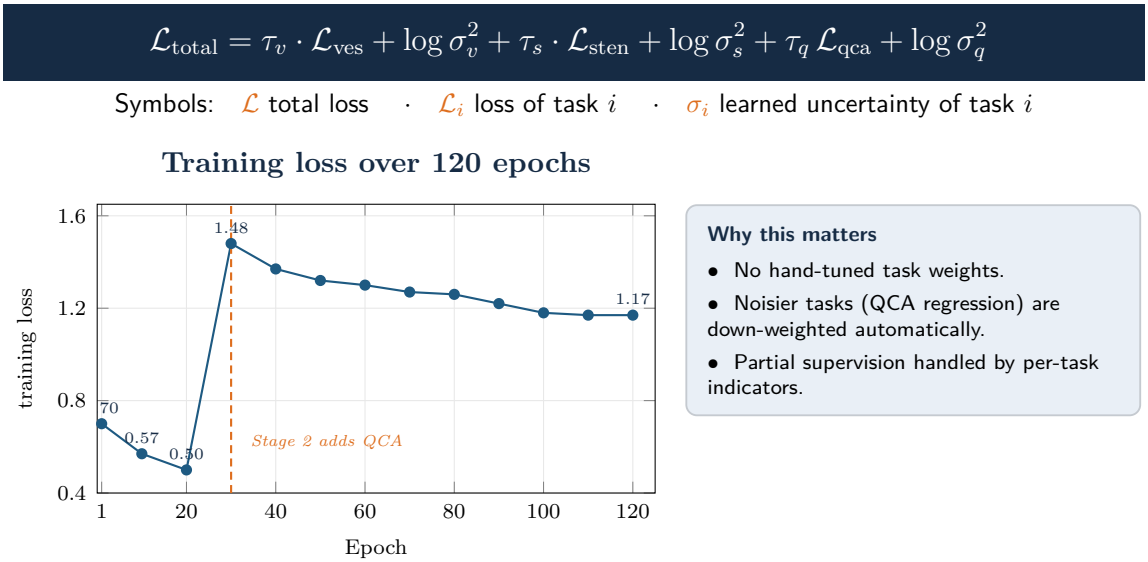


Figure 3.23: Homoscedastic uncertainty weighting of the three task losses (Equation (3.26)). The total training loss across 120 epochs falls during the segmentation warm-up (epochs 1–25), then steps up sharply near epoch 30 when Stage 2 activates the QCA term, before declining steadily through the long fine-tune. The learnable log-variances mean no task weights are hand-tuned, the noisier QCA regression is down-weighted automatically, and partial supervision is handled by the per-task indicator variables.

contribution of samples that do not carry the corresponding label.

Vessel loss. A weighted combination of Dice and BCE (Sections 1.6.5 and 1.6.5) with equal mixing weights:

$$\mathcal{L}_{\text{ves}} \;=\; 0.5\,\mathcal{L}_{\text{Dice}}(v_{\text{logits}}, v_{\text{tgt}}) + 0.5\,\mathcal{L}_{\text{BCE}}(v_{\text{logits}}, v_{\text{tgt}}). \tag{3.27}$$

The per-sample weight w_b^v is applied at the pixel level: for the BCE component, the per-pixel cross-entropy is multiplied by w_b^v before being averaged over the valid pixel count; for the Dice component, both the prediction and the target are pre-multiplied by w_b^v before computing the intersection and union. This is equivalent to giving manual masks unit weight, U-Net teacher pseudo-labels half weight, and excluding samples without any vessel signal entirely. (Flow 2 uses no centerline-Dice term; a topology-preserving clDice regulariser is introduced only in Flow 3, Section 3.11.)

Stenosis loss. A weighted sum of Focal Tversky and Sobel boundary losses (Sections 1.6.5 and 1.6.5):

$$\mathcal{L}_{\text{sten}} \;=\; 0.7\,\mathcal{L}_{\text{FT}}(s_{\text{logits}}, s_{\text{tgt}}) + 0.3\,\mathcal{L}_{\text{Bnd}}(s_{\text{logits}}, s_{\text{tgt}}). \tag{3.28}$$

The Focal Tversky loss uses $\alpha = 0.7$, $\beta = 0.3$ and $\gamma = 4/3$. These are not tuned freely but follow the values recommended by Abraham and Khan [57], who introduced the Focal

Tversky loss precisely for small-lesion medical segmentation. The asymmetry $\alpha > \beta$ (with $\alpha + \beta = 1$) makes the Tversky index penalise false negatives more heavily than false positives, which is the desired behaviour for stenosis localisation: a stenotic lesion occupies under 1% of the image, so missing it (a false negative) is far costlier than a small over-segmentation. The focusing exponent $\gamma = 4/3 > 1$ further down-weights easy, already well-segmented examples and concentrates the gradient on hard, low-overlap cases, mirroring the role of the focusing term in focal loss [10]. These settings were retained after they gave stable convergence on the validation split; a full ablation over (α, β, γ) is left as future work. Both components apply the per-sample weight w_b^s at the pixel level. If no sample in the current mini-batch carries a valid stenosis mask ($\sum_b w_b^s = 0$), $\mathcal{L}_{\text{sten}}$ is set to zero and excluded from the total via the indicator $\nVdash_s$ in Equation (3.26).

QCA loss. A masked Smooth L1 (Huber) loss (Section 1.6.5) with $\beta = 0.1$:

$$\mathcal{L}_{\text{qca}} \;=\; \frac{\sum_b w_b^q \cdot \mathcal{L}_{\text{Huber}}(\hat{q}_b, q_b)}{2 \sum_b w_b^q + \varepsilon}, \tag{3.29}$$

where the factor of two in the denominator accounts for the two regression targets (%DS and lesion length, both expressed in their normalised $[0, 1]$ form). The validity mask ensures that patients without QCA tags do not contribute spurious zeros, and the same indicator mechanism $\nVdash_q$ excludes the term entirely when no sample in the mini-batch carries QCA ground truth.

3.8 Training Strategy

3.8.1 Stratified Patient-Level Split and the Role of the Held-Out Set

Patients are partitioned into a training set (85%) and a held-out set (15%) by label-balanced stratification: patients are first grouped according to which combination of labels they carry (both stenosis mask and QCA, stenosis only, QCA only, neither), and a fraction is drawn proportionally from each group. This guarantees that the held-out set contains representative examples of every label configuration; without this stratification, multi-task metrics would be undefined for entire sub-populations. The split is performed once with a fixed seed and stored in a JSON file for reproducibility.

On the use of "validation" versus "test". A clarification of terminology is in order, because it affects how the reported numbers should be read. In the strict three-way protocol, a model is tuned on a training set, its hyper-parameters and checkpoints are selected on a validation set, and its final performance is reported once on an untouched

test set. The present study operates under a single-centre, small-data constraint (200 patients): a three-way split would leave each partition too small to yield stable metrics. We therefore use a single held-out partition that plays a *dual role*. Checkpoint selection during the staged curriculum (Section 3.8.4) reads metrics on this partition, so in the strict sense it is a *validation* set; the same partition is then the one on which all reported metrics are computed, so the figures in Section 3.10 should be interpreted as *validation-set* performance rather than as a fully independent test-set estimate. We use the word "validation" consistently for this partition throughout the chapter to avoid implying an independence that the data volume does not permit. A genuine external test set — ideally from a second centre — is identified as the most important item of future work (Section 3.15); the three-fold cross-validation implemented for Flow 3 (Section 3.11.7) is a first step in that direction, reducing the dependence of the reported numbers on a single split.

3.8.2 Data Augmentation Pipeline

During training, each batch is augmented with the following transformations, applied jointly to the image and both masks: horizontal flip ($p = 0.5$), vertical flip ($p = 0.3$), rotation $\pm 20°$ ($p = 0.5$, with constant-value border padding), random brightness/contrast ± 0.15 ($p = 0.5$), Gaussian noise ($p = 0.3$, variance range $[10, 50]$), elastic deformation ($p = 0.2$, $\alpha = 30$, $\sigma = 5$), and grid distortion ($p = 0.2$, five steps, distortion limit 0.2). Validation uses only the resize to 512512 and tensor conversion. The implementation relies on the `Albumentations` image-augmentation library [20]; the `DataLoader` uses two workers per loader with persistent workers and pinned memory.

3.8.3 Optimiser and Discriminative Learning Rates

The optimiser is AdamW [14] with weight decay 10^{-4} and three parameter groups: the pretrained MaxViT encoder, the freshly initialised heads (both decoders, both segmentation 11 output convolutions, the ROI pooling projection and the QCA MLP), and the loss-module parameters (the three log-variances). Discriminative learning rates are used: the encoder is trained at half the base learning rate of the heads ($\eta_{\text{enc}} = 0.5\,\eta$), encouraging the pretrained features to adapt gently while the new heads converge faster. Mixed-precision training (Section 1.6.8) is used throughout, with gradient clipping at $\|\nabla\| \leq 1.0$ applied to both the model parameters and the loss-module parameters in order to stabilise the early epochs.

3.8.4 Three-Stage Training Curriculum

A curriculum-style staged training is used to manage the differing signal qualities of the three tasks. The configuration is summarised in Table 3.5.

Table 3.5: Three-stage training schedule used for GCT-Net (Flow 2). Batch size is 4 on a single Tesla T4 GPU; η_{head} denotes the head learning rate, with the encoder at half this rate. A hard stop-gradient is applied at the stenosis-to-QCA junction throughout all three stages.

Stage	Goal	Trainable parameters	Epochs	Learning-rate schedule
1	Warm-up of segmentation; QCA loss masked out by overriding `qca_valid`	Encoder + both decoders + segmentation heads	25	OneCycleLR with peak $\eta_{\text{head}} = 210^{-4}$, `pct_start` = 0.1, `div_factor` = 10, `final_div_factor` = 50
2	Joint optimisation with QCA active	All parameters including the three log-variances	35	CosineAnnealingLR over 35`steps_per_epoch` steps, $\eta_{\text{head}} = 110^{-4}$
3	Long fine-tune at low LR	All parameters	60	CosineAnnealingWarmRestarts with $T_0 = 20$`steps_per_epoch` and $T_{\text{mult}} = 2$, $\eta_{\text{head}} = 310^{-5}$

In Stage 1, the QCA head is excluded from the loss by overriding `qca_valid` to zero in every batch (the indicator $\mathbb{1}_q$ in Equation (3.26) therefore evaluates to zero). The encoder is thus free to learn features useful for vessels and stenoses before being asked to support the regression head. The best checkpoint of Stage 1 is selected by the maximum held-out stenosis Dice. Stage 2 activates the QCA loss; the three log-variances now receive gradients and the encoder is fine-tuned on the joint objective. Stage 2 selection uses a composite QCA score

$$S_{\text{S2}} \;=\; \text{MAE}_{\%DS} + 5 \cdot \text{MAE}_{\ell}, \tag{3.30}$$

where $\text{MAE}_{\%DS}$ is in percentage points and MAE_{ℓ} in millimetres; the factor of five normalises the two quantities to comparable scales. Stage 3 refines all three outputs over a longer horizon; the best checkpoint is selected by

$$S_{\text{S3}} \;=\; \text{MAE}_{\%DS} - 50 \cdot \text{Dice}_{\text{sten}}, \tag{3.31}$$

which rewards both low QCA error and high stenosis Dice in a single scalar. The factor of 50 matches the typical numerical range of the percentage MAE, so neither term dominates a priori.

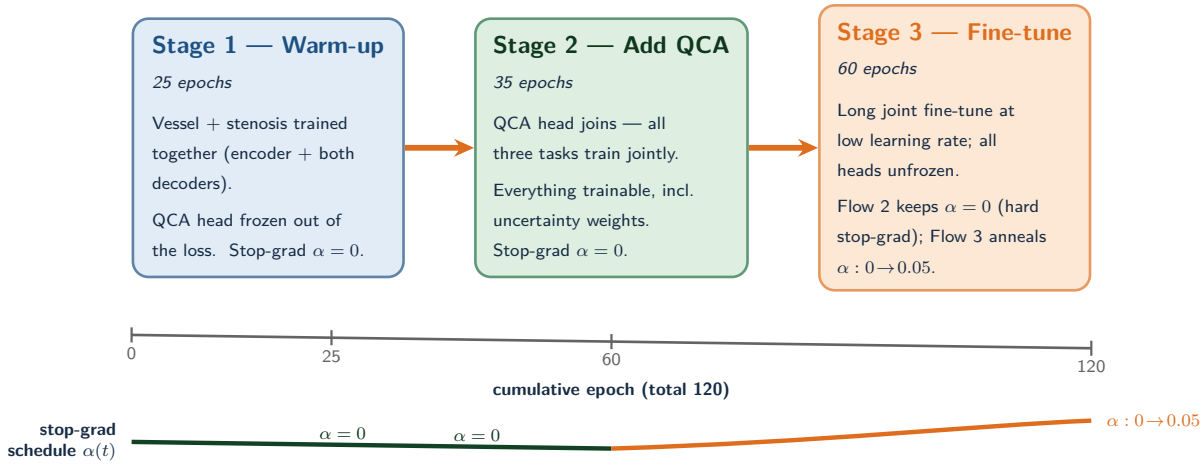


Figure 3.24: The three-stage training curriculum of GCT-Net (Flow 2). Stage 1 (25 epochs) warms up the two segmentation tasks with the QCA head frozen out of the loss; Stage 2 (35 epochs) activates the QCA head so all three tasks train jointly, with the uncertainty weights now learnable; Stage 3 (60 epochs) is a long joint fine-tune at a low learning rate. The lower strip shows the stop-gradient schedule $\alpha(t)$: the gradient gate stays at $\alpha = 0$ (hard separation) throughout Stages 1 and 2 in Flow 2; the annealed ramp $\alpha : 0 \to 0.05$ shown for context is introduced only in Flow 3 (Section 3.11). The three stages span 120 cumulative epochs, and the step-up in the training loss near epoch 30 (Figure 3.23) marks the Stage 2 activation of the QCA term.

3.9 Inference Pipeline

At inference time, the same preprocessing pipeline is applied to a new DICOM frame. The resulting tensor is passed to GCT-Net, producing simultaneously a vessel logit map, a stenosis logit map, and the two normalised QCA values $(\widehat{\%DS}, \hat{\ell})$. The raw outputs are refined by test-time augmentation and an adaptive thresholding step.

Four-flip test-time augmentation. The model is evaluated on four geometric variants of the input — the original frame, its horizontal flip, its vertical flip, and the combined horizontal–vertical flip. Each output segmentation map is sigmoid-activated and geometrically inverted back to the original orientation; the four resulting probability maps are averaged, as are the four QCA outputs. This TTA reduces output variance at the cost of a four-fold inference time.

Adaptive thresholding of the stenosis map. A fixed binarisation threshold (e.g., 0.5) is poorly suited to the stenosis map because the predicted probabilities often peak at much lower values for small or ambiguous lesions, in which case a fixed threshold would yield an empty mask. Let $\rho = \max_{u,v} p_s(u, v)$ be the maximum stenosis probability after

TTA, and let $\tau_{\text{base}} = 0.30$. The threshold actually applied is

$$\tau \;=\; \begin{cases} \max(0.05,\; 0.4\,\rho) & \text{if } \rho < 0.20, \\ \min(\tau_{\text{base}},\; 0.5\,\rho) & \text{otherwise.} \end{cases} \tag{3.32}$$

where τ is the threshold actually applied to binarise the stenosis probability map, ρ is the post-TTA peak stenosis probability defined above, and $\tau_{\text{base}} = 0.30$ is the nominal threshold. When the peak confidence is low ($\rho < 0.20$), the threshold is lowered to $0.4\,\rho$ (with a floor of 0.05) so that a faint but genuine lesion is not erased; otherwise the threshold is set to $0.5\,\rho$ but capped at τ_{base}, so a confident prediction is never thresholded more aggressively than the nominal value. This keeps small, low-probability lesions detectable while preventing spurious activations on high-confidence frames.

The binarised stenosis mask, the vessel probability map, the input angiogram, an overlay of manual ground truth (when present, in green) against the prediction (in red), and a bar chart of predicted versus DICOM-reported $\%DS$ and lesion length together form the per-patient visualisation grid. (The additional deployment-time refinements — the teacher-based anatomical vessel filter, the vessel-gating of the stenosis map, the no-stenosis gate, and the conformal prediction interval — are introduced in Flow 3, Section 3.11, and used by the web application of Section 3.14.)

Validation uses the four-flip TTA described above. Vessel Dice is computed at threshold $\tau_{\text{ves}} = 0.50$ over samples carrying a manual vessel mask only; in the present cohort no manual vessel masks were withheld for validation ($n = 0$), so the vessel head is supervised entirely by the teacher pseudo-labels and the validation vessel Dice is undefined. Stenosis Dice is computed over samples carrying a manual stenosis mask, with the adaptive threshold described above. QCA metrics are reported as the mean absolute error of the de-normalised predictions ($\text{MAE}_{\%DS}$ in percentage points after multiplying the network output by 100, MAE_{ℓ} in millimetres after multiplying by 50). Samples lacking the relevant ground truth are excluded from the corresponding metric only.

3.10 Experimental Results

All metrics in this section are computed on the held-out validation partition defined in Section 3.8.1; as discussed there, this partition is used both for checkpoint selection and for the reported figures, and the numbers should be read accordingly.

3.10.1 Training Dynamics

Table 3.6 reports the best per-stage validation metrics over the three-stage curriculum. Because no manual vessel masks were withheld for validation in this cohort, the validation

vessel Dice is undefined throughout and is not reported. Stage 1 is dominated by the two segmentation losses and lifts the stenosis Dice from near-zero to 0.38 within 25 epochs (QCA frozen out of the loss). Stage 2 activates the QCA head and pulls the stenosis Dice up to 0.58 while reducing the QCA $\%DS$ MAE from ~ 20 to ~ 11 pp and the length MAE to ~ 4.6 mm. Stage 3 fine-tunes all three outputs at low learning rate; the stenosis Dice peaks at 0.66 and the $\%DS$ MAE bottoms at 9.80 pp. The best checkpoint of Stage 3 (selected by the composite $\text{MAE}_{\%DS} - 50 \cdot \text{Dice}_{\text{sten}}$) is the model used for all subsequent results; because that composite balances QCA error against stenosis overlap, the selected checkpoint reports a stenosis Dice of 0.610 rather than the single-epoch peak of 0.657, trading a small amount of overlap for the lowest joint error.

Table 3.6: Best validation metrics achieved during each training stage of GCT-Net (Flow 2) on fold 0. Vessel Dice is undefined (no withheld manual vessel masks).

Stage	Epochs	Best $\text{Dice}_{\text{sten}}$	Best $\text{MAE}_{\%DS}$	Best MAE_{ℓ}
1 (segmentation warm-up)	25	0.377	—	—
2 (joint optimisation)	35	0.577	10.74 pp	4.59 mm
3 (long fine-tune)	60	0.657	9.80 pp	4.04 mm

The homoscedastic log-variances drift from their initial $(0, 0, 1)$ toward roughly $(\log \sigma_v^2, \log \sigma_s^2, \log \sigma_q^2) \approx (-0.08, -0.08, +1.0)$ over training, corresponding to relative effective weights of approximately $(\tau_v, \tau_s, \tau_q) \approx (1.08, 1.08, 0.37)$. As predicted by Section 1.6.5, the QCA targets — parsed from DICOM private tags and inherently noisier than pixel masks drawn by clinical staff — automatically receive a smaller effective weight, without manual tuning of loss coefficients.

The full per-epoch training dynamics of all three stages are shown in Figure 3.25, plotted directly from the training log.

3.10.2 Segmentation Results

Table 3.7 reports the validation results on the patient-level held-out set ($n = 21$ patients, fold 0) for the two segmentation tasks. No manual vessel masks were withheld in this cohort ($n = 0$), so the vessel head is supervised entirely by the U-Net teacher pseudo-labels and no validation vessel Dice is reported. The stenosis Dice of 0.610 ± 0.234 is markedly higher than the prevalence-weighted area would suggest: a stenotic lesion typically occupies fewer than 1% of the image pixels, so any non-trivial Dice on this regime is a non-trivial result, and it improves substantially over the 0.345 obtained by the separately trained Flow 1 stenosis model.

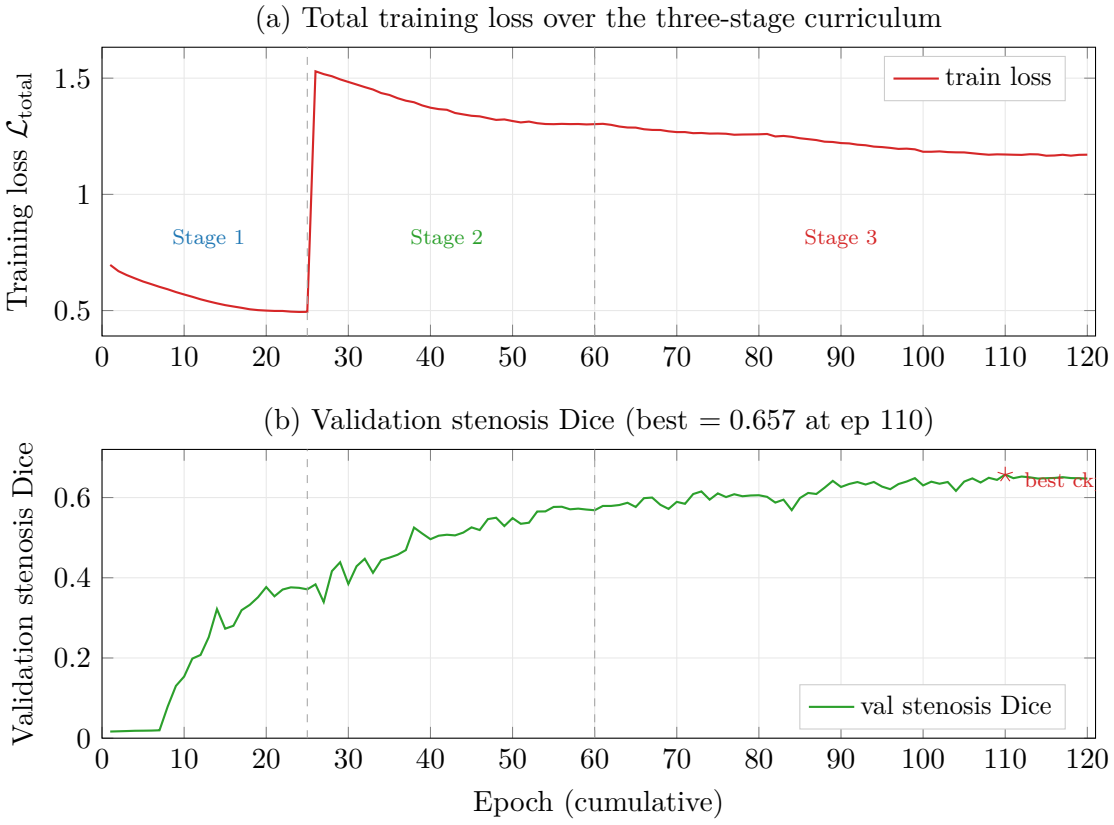


Figure 3.25: Training dynamics of GCT-Net (Flow 2) over the three-stage curriculum (Table 3.5), plotted from the per-epoch training log. Vertical dashed lines mark the stage boundaries (Stage 1: warm-up, epochs 1–25; Stage 2: joint optimisation with QCA active, epochs 26–60; Stage 3: long fine-tune, epochs 61–120). (a) The total loss drops sharply in Stage 1, jumps when the QCA term is activated at the start of Stage 2, then declines steadily through Stage 3. (b) The validation stenosis Dice rises from near zero to 0.377 by the end of Stage 1, climbs to 0.577 during Stage 2, and peaks at 0.657 in Stage 3, where the best checkpoint (star) is selected by $\text{MAE}_{\%DS} - 50 \cdot \text{Dice}_{\text{sten}}$.

Table 3.7: Segmentation results of GCT-Net (Flow 2) on the validation split (fold 0). Stenosis Dice is reported as mean ± standard deviation under four-flip TTA and the adaptive threshold.

Task	Dice (mean)	n (val)
Vessel segmentation (manual masks only)	n/a	0
Stenosis localisation	0.610 ± 0.234	21

3.10.3 QCA Quantification Results

Table 3.8 reports the QCA agreement on the 21 validation patients carrying valid QCA tags, after de-normalising the predictions. The $\%DS$ MAE of 11.73 percentage points and the lesion-length MAE of 4.51 mm are obtained under four-flip TTA from the best Stage 3 checkpoint. Both errors are substantially lower than the corresponding Flow 1 figures (18.9% and 5.7 mm), confirming that conditioning the QCA regression on the predicted lesion through lesion-aware ROI pooling, within a jointly optimised network, recovers accuracy that the cascaded pipeline loses.

Table 3.8: QCA quantification results of GCT-Net (Flow 2) on the validation split (fold 0, $n = 21$ patients with valid DICOM QCA tags), reported as mean absolute error of the de-normalised predictions.

Quantity	MAE	n
Stenosis $\%DS$	11.73 pp	21
Lesion length ℓ	4.51 mm	21

The evolution of both QCA errors across the joint and fine-tune stages is shown in Figure 3.26.

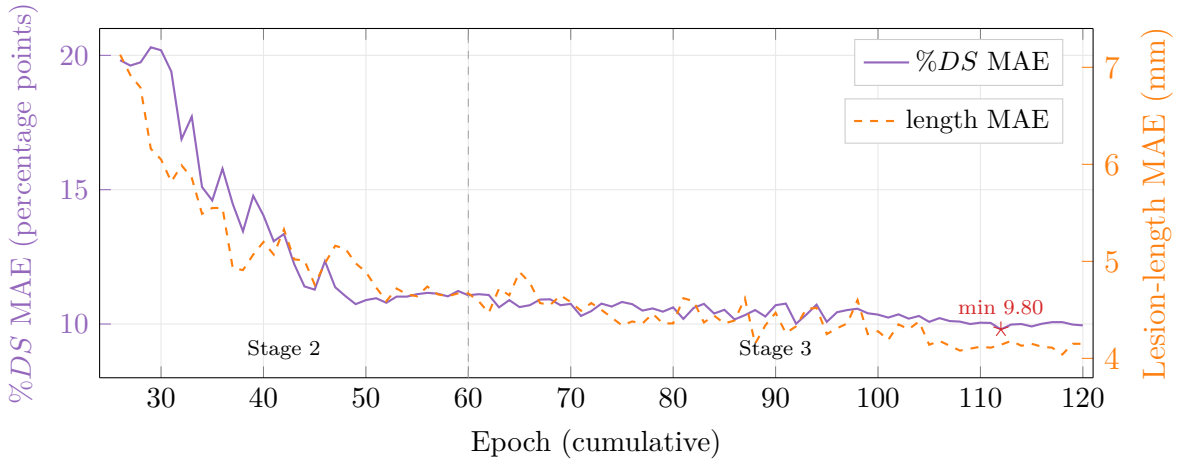


Figure 3.26: QCA regression error of GCT-Net (Flow 2) on the validation split during Stages 2 and 3 (the QCA loss is masked out in Stage 1, so no QCA metric is defined there). The percentage-diameter-stenosis MAE (left axis, purple) falls from ≈ 20 pp at the start of Stage 2 to its minimum of 9.80 pp; the lesion-length MAE (right axis, orange dashed) falls from ≈ 7 mm to ≈ 4.0 mm. The reported best-checkpoint values are 11.73 pp and 4.51 mm under four-flip TTA (Table 3.8).

3.10.4 Qualitative Results

For each validation patient, an inference visualisation grid is produced containing six panels (Figure 3.27): the input angiogram; the predicted vessel probability map; the

manual stenosis mask (when available, shown in green); the predicted stenosis mask after adaptive thresholding (shown in red); an overlay placing both masks on the original angiogram for direct comparison; and a bar chart showing the predicted $\%DS$ and ℓ alongside the DICOM-reported values. Patients are sorted by stenosis Dice from highest to lowest, so that the top of the figure corresponds to easy cases and the bottom to failure cases. Visual inspection confirms three patterns: (i) on patients with a clearly delineated, well-opacified lesion (top of the figure), the predicted and ground-truth stenosis masks overlap substantially and the QCA error is small; (ii) on patients with diffuse disease (middle), the predicted mask captures the same anatomical region but is shifted by a few pixels, which lowers the Dice without greatly harming the QCA estimate; (iii) on the worst cases (bottom), the predicted lesion is on a different vessel branch from the ground-truth lesion, indicating that branch-level disambiguation rather than pixel-level localisation is the dominant remaining failure mode.

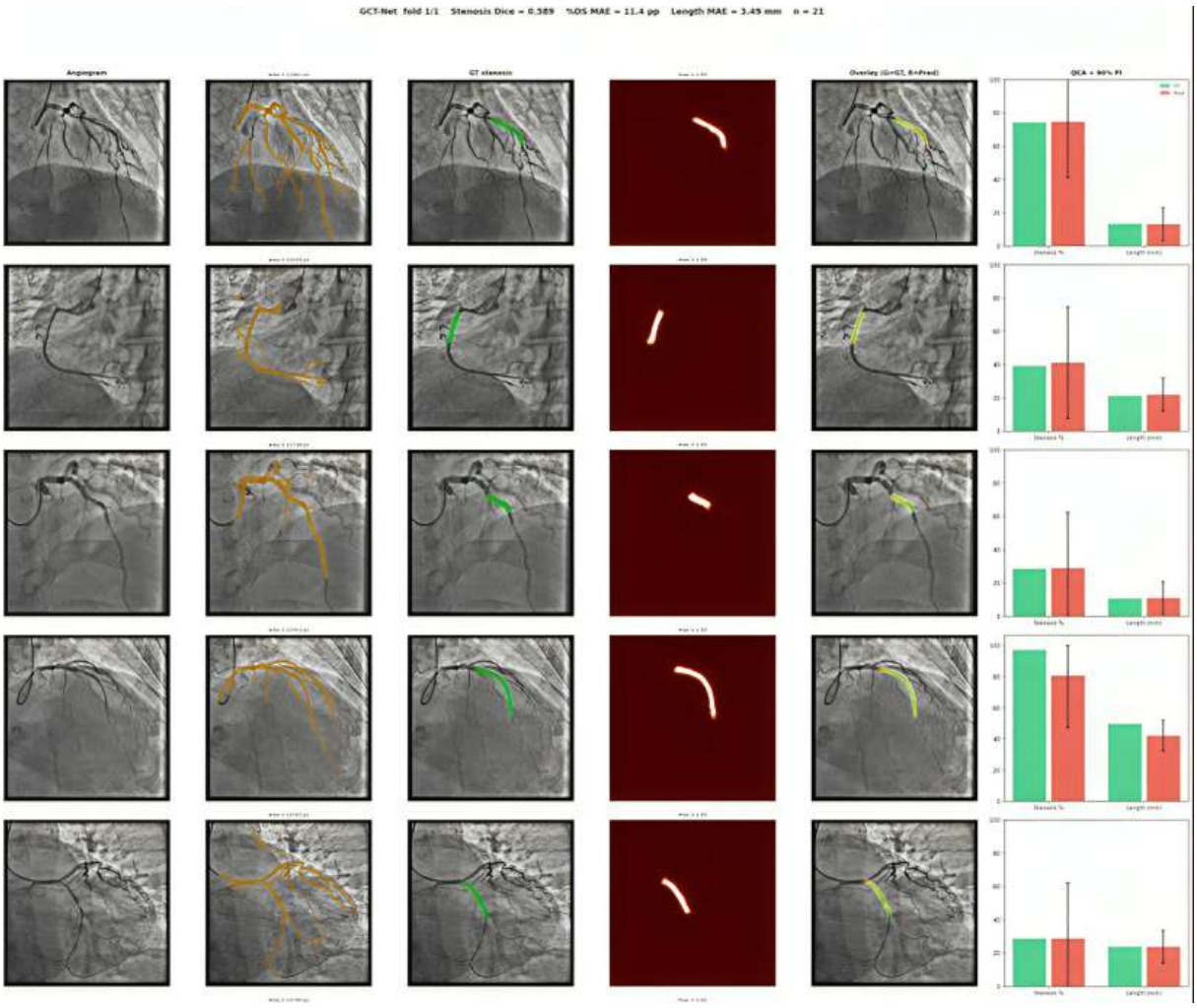


Figure 3.27: Qualitative validation results of GCT-Net (Flow 2) on fold 0 ($n = 21$ patients, sorted by stenosis Dice). Each row shows, from left to right: input angiogram, predicted vessel probability, manual stenosis ground truth, predicted stenosis after adaptive thresholding, the overlay (green for ground truth, red for prediction), and a bar chart of predicted versus DICOM-reported $\%DS$ and lesion length.

3.10.5 Limitations of the Single-Source Model

Three limitations of the unified Flow 2 model motivate the multi-source extension of Flow 3. They are stated directly below, each with the schematic that makes the issue concrete.

1 — Asymmetric coupling. The coupling between the stenosis and QCA tasks is one-directional. Because the stop-gradient blocks the QCA gradient from reaching the stenosis decoder, gradients flow one way only (stenosis → QCA): the QCA head rides on the predicted stenosis map, so a small localisation error in that map misleads the regression with no path to recover. The very mechanism that protects the stenosis map (Contribution 2) also prevents the QCA head from ever correcting it.

Gradients flow one way only. QCA rides on the stenosis map, so a small lesion error misleads it — with no path to recover.

Figure 3.28: Limitation 1 — asymmetric coupling. The stop-gradient ($\alpha = 0$) makes the stenosis→QCA dependence one-directional, so errors in the stenosis map propagate into the QCA estimate but the QCA loss cannot help fix the map.

2 — Brittle curriculum. The staged curriculum is hand-designed rather than derived. The order of the three stages and the learning-rate schedule were chosen by hand, so a different order or schedule could reach a different optimum; the reported result therefore hinges on design choices that are not themselves optimised or fully reported.

Stage order and α-ramp hand-tuned, not derived. A different order can reach a different optimum, so results hinge on unreported choices.

Figure 3.29: Limitation 2 — brittle curriculum. The three-stage order and schedule are hand-chosen, not optimised, so the final result depends on design decisions that are not derived from the data.

3 — Non-modular design. Hard parameter sharing means a single shared encoder serves all tasks, which is efficient but rigid: adding a fourth task would force the whole encoder to be retrained from scratch rather than extended as a plug-in module. Flow 3

addresses this point directly by adding new supervision and modalities on top of the same backbone.

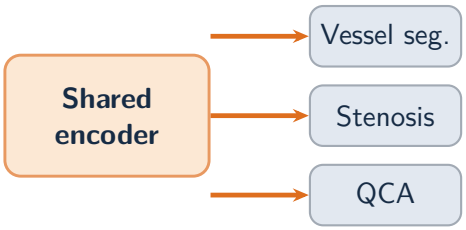


Figure 3.30: Limitation 3 — non-modular design. A single shared encoder feeds all heads, so introducing a fourth task forces a full retrain of the encoder rather than a modular add-on. Flow 3 (Section 3.11) tackles this by building new supervision and modalities onto the same backbone.

3.11 Flow 3 — Multi-Source / Multi-Task Extension: GCT-Net

The third flow extends the single-source GCT-Net into a multi-source, multi-task model, GCT-Net, which ingests the same information a cardiologist consults: the angiogram, the resting electrocardiogram (ECG), and the procedural report. The three original image tasks are preserved so that the Flow 2 behaviour can be recovered exactly by disabling the new modality and supervision flags; the extension adds a non-visual ECG branch, four report-derived supervision heads, a topology-preserving vessel loss, an annealable stenosis–QCA coupling, and a split-conformal predictor on the QCA outputs.

3.11.1 Heterogeneous Inputs

Each patient is represented by three aligned sources. The *angiogram* is the single contrast-opacified frame selected by the automatic frame selector (Section 3.6.3), the spatial substrate for vessel and stenosis localisation. The *ECG* is a 13-dimensional tabular feature vector read from a structured `Clinical12_Features.csv` produced by the open-source NeuroKit2 library [4]: twelve numeric clinical features (heart rate, RR, PR, QRS, QT, QTc, P/R/T amplitudes, ST deviation, heart-rate variability, P-wave duration) plus one quality flag that gates the auxiliary losses. The *procedural report* is the cath-lab `compte-rendu.docx`, parsed once by a large-language-model extractor (Gemini 3.1 Pro) into a structured table and consumed as feature fields rather than raw text: the indication (STABLE / NSTEMI / STEMI / OTHER), the treated arteries (a multi-label over LAD, LCx, RCA, LM), the per-artery stent dimensions, and the binary procedural success.

3.11.2 Modality-Specific Encoders and FiLM Fusion

The vision encoder re-uses the Flow 2 MaxViT-tiny backbone. The ECG vector is encoded by a lightweight tabular MLP — an input `BatchNorm` followed by two hidden layers ($13 \to 128 \to 128$) with GELU activations and dropout and a final projection to a 256-dimensional embedding with `LayerNorm` and SiLU — which normalises the disparate scales of the ECG features without manual scaling and is robust to missing fields through the quality flag.

The two modalities are combined by Feature-wise Linear Modulation [34] (FiLM) at the image bottleneck. The ECG embedding produces a per-channel scale γ and shift β that modulate the image feature map x as

$$\mathrm{FiLM}(x) \;=\; x \cdot \big(1 + \mathrm{gate} \cdot \gamma\big) + \mathrm{gate} \cdot \beta, \tag{3.33}$$

where gate is the per-sample `ecg_valid` flag. The projections generating γ and β are initialised to zero, so that before training the FiLM block is the identity and the image backbone is unperturbed early on; when the ECG is missing the gate is zero and the layer degrades gracefully to the identity. This is a far lighter conditioning mechanism than full cross-attention, adding only two linear projections at the bottleneck.

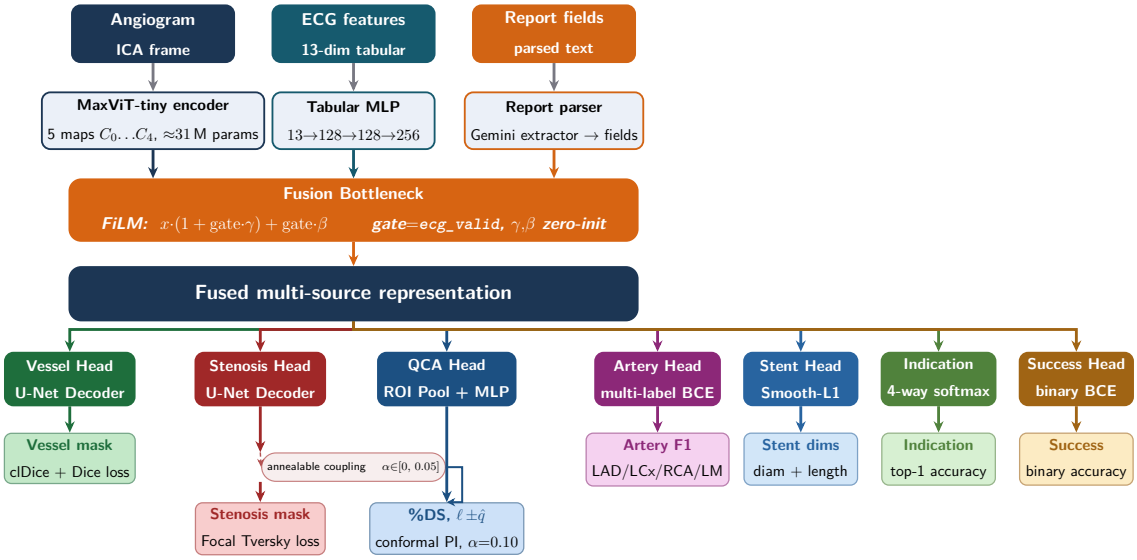


Figure 3.31: Architecture of GCT-Net (Flow 3). Three input modalities are encoded separately: the angiogram by the shared MaxViT-tiny backbone (re-used from Flow 2), the resting ECG by a tabular MLP, and the parsed report fields as structured features. The encoders are fused at a FiLM bottleneck gated by `ecg_valid`. Three image heads (vessel, stenosis, QCA) and four report-derived heads (artery classification, stent regression, indication classification, procedural success) operate on the fused representation. An annealable coupling $\alpha\in[0, 0.05]$ replaces the hard stop-gradient of Flow 2. A clDice term is added to the vessel loss and the QCA outputs are wrapped in a split-conformal predictor (α=0.10).

3.11.3 Novel Contributions

Three methodological contributions distinguish Flow 3 from the single-source model. As in Flow 2, each is presented as the problem it solves, the mechanism that solves it (with its governing equation), and the consequence it delivers.

Contribution 1 — FiLM Fusion: Conditioning Vision on ECG

The problem. The ECG arrives as a non-spatial 13-dimensional vector, while the image encoder produces a spatial feature map. Injecting the former into the latter without resorting to heavy cross-attention machinery — which would add many parameters and risk overfitting on a small cohort — is the core fusion problem.

How it works. Feature-wise Linear Modulation (FiLM) [34] conditions the vision features on the ECG with only two linear projections. The ECG MLP embedding produces a per-channel scale γ and shift β, which modulate the image bottleneck feature map x as

$$\mathrm{FiLM}(x) \;=\; x \cdot \big(1 + \mathrm{gate} \cdot \gamma\big) + \mathrm{gate} \cdot \beta, \tag{3.34}$$

where each symbol carries a deliberate role. x is the image feature map from the MaxViT bottleneck; γ and β are the channel-wise scale and shift generated from the ECG embedding; and gate is the per-sample `ecg_valid` flag, equal to 1 when an ECG is present and 0 when it is missing. Reading Equation (3.34) term by term: the factor $(1 + \mathrm{gate} \cdot \gamma)$ multiplicatively rescales each feature channel, and the additive $\mathrm{gate} \cdot \beta$ shifts it. The $+1$ inside the scale factor is what makes the operation default to the identity — when $\gamma = 0$ the scale is exactly 1. The projections generating γ and β are initialised to zero, so at the start of training $\gamma = \beta = 0$ and $\mathrm{FiLM}(x) = x$: the image backbone is unperturbed and only gradually learns to let the ECG modulate it. When the ECG is absent, gate $= 0$ collapses both terms and the layer again reduces to the identity $\mathrm{FiLM}(x) = x$.

Why it matters. FiLM has trivial parameter cost compared with cross-attention; the zero-initialised γ, β keep the pretrained vision backbone stable early in training; and the gate gives graceful degradation — a patient with no ECG simply receives the unmodulated image features rather than a corrupted representation.

Contribution 2 — Multi-Source Supervision: Reports as Labels

The problem. Each case ships a free-text procedural report that is rich in clinical detail but, in raw narrative form, cannot be used as a training label. The challenge is to convert that narrative into supervised signals without any additional manual annotation.

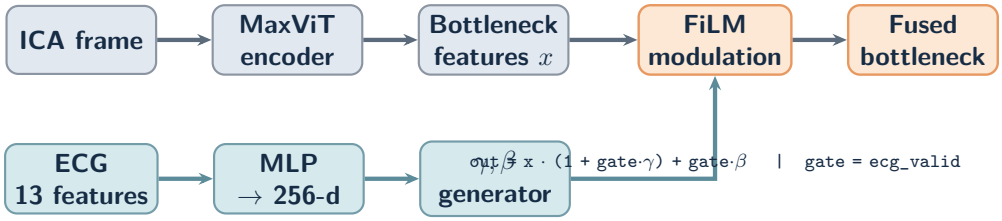


Figure 3.32: Contribution 1 — FiLM fusion. The ICA frame is encoded by the MaxViT backbone into bottleneck features x (top), while the 13-feature ECG vector is encoded by an MLP that produces the FiLM scale γ and shift β (bottom). FiLM modulates x by Equation (3.34), gated by `ecg_valid`, to yield the fused bottleneck. With γ, β zero-initialised and the gate disabling the layer when the ECG is missing, the block defaults to the identity.

How it works. The report is parsed once (by a large-language-model extractor) into structured fields, which become the targets of four auxiliary supervision heads attached to the fused representation: the treated arteries (a multi-label classification over LAD, LCx, RCA, LM), the stent dimensions (a regression over diameter and length), the indication (a four-way classification), and the procedural success (a binary classification). Each head reads the same fused patient-state representation and is supervised only when the report for that patient was parseable.

Why it matters. Free-text narratives that were previously unusable become four clinically meaningful supervised signals, so the model learns additional structure with no extra manual labelling effort beyond the one-time parsing step.

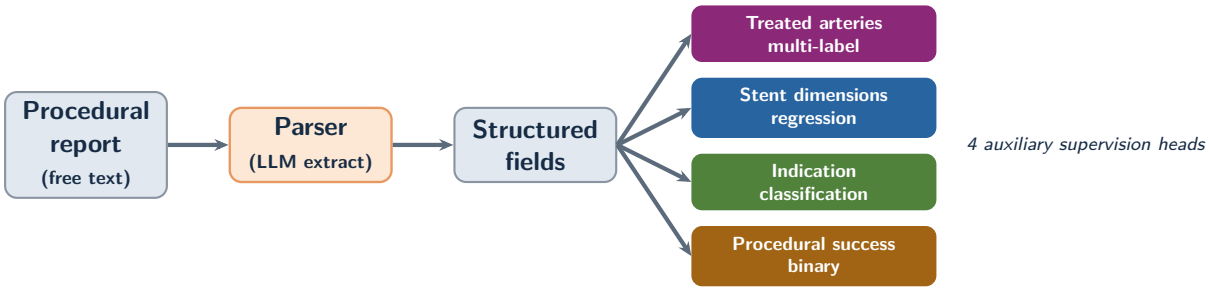


Figure 3.33: Contribution 2 — multi-source supervision. The free-text procedural report is parsed once into structured fields, which supervise four auxiliary heads on the fused representation: treated arteries (multi-label), stent dimensions (regression), indication (classification), and procedural success (binary). The narrative thus becomes four supervised signals at no extra manual labelling cost.

Contribution 3 — Structured Reporting, Balanced Loss and Robustness

The problem. Three practical demands must be met at once: the predictions must reach clinicians as a usable report; seven tasks must train in balance without one dominating; and many patients lack an ECG or a report, so training must tolerate missing

modalities.

How it works. Three mechanisms address these in turn. First, *deterministic report generation*: the predicted structured fields drive a fixed rule-based template (no language model in the loop) that renders a Word-compatible procedural report. Second, a *seven-task balanced loss*: the homoscedastic uncertainty weighting of Flow 2 is extended to seven learnable log-variances, one per task, so all seven losses are weighted automatically; this stage also activates the clDice vessel term and the annealed stenosis–QCA coupling (both detailed in Section 3.11.5). Third, *per-sample validity masking*: a quality flag zeroes the auxiliary losses for any patient whose ECG or report is missing, so those patients still drive the imaging losses without contributing spurious auxiliary gradients.

Why it matters. Deterministic templating yields reproducible, clinician-ready reports; the learned weighting balances all seven tasks without manual tuning; and validity masking keeps training robust under the partial observation that is the norm in a real single-centre archive.

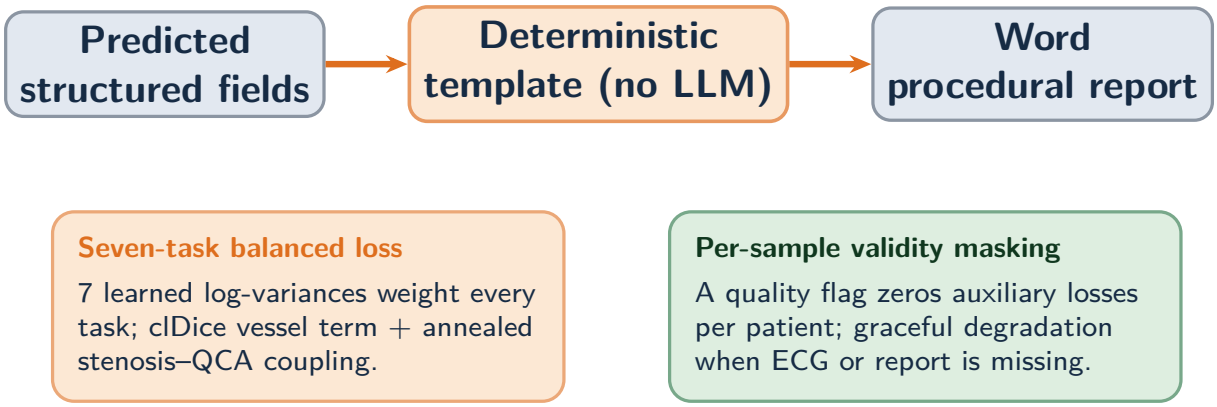


Figure 3.34: Contribution 3 — structured reporting, balanced loss and robustness. Top: the predicted structured fields drive a deterministic, no-LLM template that renders a Word procedural report. Bottom: the seven-task homoscedastic loss (left) balances all tasks automatically and carries the clDice and annealed-coupling terms, while per-sample validity masking (right) zeroes auxiliary losses for patients lacking an ECG or report, giving graceful degradation under partial supervision.

3.11.4 Auxiliary Report-Derived Heads

Four small MLP heads share the fused representation and convert the free-text report into supervised signals. The *artery head* is a four-way multi-label sigmoid classifier (LAD, LCx, RCA, LM) supervised by the treated-artery vector with a binary cross-entropy loss. The *stent head* is an eight-dimensional regression (diameter and length per artery) supervised by a Smooth-L1 loss, normalised against bounds of 6 mm diameter and 60 mm length. The

indication head is a four-way softmax classifier trained with cross-entropy. The *success head* is a binary classifier trained with BCE. All four are masked per sample by a report-validity flag, so patients without a parseable report still drive the segmentation and QCA losses without spurious auxiliary gradients.

3.11.5 Seven-Task Loss, Annealable Coupling, clDice and Conformal Calibration

Flow 3 trains seven heads at once, so its objective must (i) combine seven losses of very different type and scale, (ii) tolerate the fact that most patients carry only a subset of the seven labels, and (iii) add the two mechanisms that were deliberately held out of Flow 2 — a topology-preserving vessel term and a soft, annealed version of the stenosis–QCA stop-gradient. This subsection develops each of these in turn and closes with the distribution-free calibration applied to the QCA outputs.

The Seven-Task Homoscedastic Objective

The three image losses of Flow 2 (vessel, stenosis, QCA) are retained and four report-derived auxiliary losses (treated arteries, stent dimensions, indication, procedural success) are added, for seven tasks in total. They are balanced by the same homoscedastic-uncertainty mechanism introduced in Flow 2 (Section 3.7), now extended from three to seven learnable log-variances. Writing the per-task precision as $\tau_t = \exp(-\log \sigma_t^2) = 1/\sigma_t^2$, the total objective is

$$\mathcal{L}_{\text{total}} \;=\; \sum_{t=1}^{7} \mathbb{1}_t \big(\tau_t \, \mathcal{L}_t + \log \sigma_t^2 \big), \tag{3.35}$$

where $\mathcal{L}_t$ is the loss of task t, $\log \sigma_t^2$ is its learnable log-variance, and $\mathbb{1}_t$ is a batch-level indicator that equals one only when at least one sample in the mini-batch carries a valid label for task t. The structure of Equation (3.35) carries three ideas at once. The precision τ_t scales each loss, so a task the network finds intrinsically noisy (large σ_t) is automatically down-weighted; the regulariser $\log \sigma_t^2$ prevents the degenerate solution of sending every variance to zero; and the indicator $\mathbb{1}_t$ implements *partial supervision* — when a task has no signal in the batch, both its data term and its log-variance term are excluded so that the absent label produces no spurious gradient on σ_t. The seven log-variances are initialised to $(0, 0, 1, 0.5, 1, 0.5, 0.5)$ — the first two (0, giving $\tau = 1$) for the two segmentation tasks, a larger value (1) for the inherently noisier QCA regression, and intermediate values for the four auxiliary report heads — biasing early training toward the cleaner segmentation signals while letting the data re-balance the weights thereafter.

The Seven Task-Specific Losses

Each task contributes a loss matched to its output type:

- **Vessel (segmentation).** A weighted sum of Dice and binary cross-entropy, augmented in Flow 3 by the topology-preserving clDice term described in Section 3.11.5. Dice and BCE reward pixel overlap; clDice additionally rewards an unbroken vessel skeleton.
- **Stenosis (segmentation).** A weighted sum of Focal Tversky and a Sobel boundary loss, with the Focal Tversky parameters $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 4/3$ carried over unchanged from Flow 2. The asymmetry $\alpha > \beta$ penalises missed lesion pixels more heavily than false alarms, which is appropriate because the stenosis occupies under 1% of the frame.
- **QCA (regression).** A masked Smooth-L1 (Huber) loss over the two normalised targets ($\%DS$ and lesion length), evaluated only on QCA-valid samples.
- **Treated arteries (multi-label classification).** A four-way multi-label sigmoid head (LAD, LCx, RCA, LM) trained with binary cross-entropy against the treated-artery vector.
- **Stent dimensions (regression).** An eight-dimensional Smooth-L1 regression (diameter and length per artery), normalised against bounds of 6 mm diameter and 60 mm length.
- **Indication (classification).** A four-way softmax head trained with cross-entropy.
- **Procedural success (classification).** A binary head trained with cross-entropy.

The four auxiliary losses are each gated by a per-sample report-validity flag, the per-task analogue of the indicator $\mathbb{1}_t$ in Equation (3.35), so a patient with no parseable report still contributes its imaging losses without injecting auxiliary gradients.

Topology-Preserving Vessel Loss (clDice)

A pixel-overlap loss such as Dice treats every misclassified pixel equally, so a model can score well while still breaking a thin vessel into disconnected fragments — a failure that is visually and clinically significant even though it costs few pixels. Flow 3 therefore adds a centerline-Dice (clDice) [55] term to the vessel loss, gated to the patients that carry a manual vessel mask. clDice is computed not on the masks themselves but on their morphological *skeletons*. Let S_P be the skeleton of the predicted vessel mask and S_T the skeleton of the ground-truth mask. Two quantities are defined: the topology precision T_{prec}, the fraction of the predicted skeleton S_P that falls inside the ground-truth mask,

and the topology sensitivity T_{sens}, the fraction of the ground-truth skeleton S_T that falls inside the predicted mask. clDice is their harmonic mean,

$$\text{clDice} \;=\; 2\,\frac{T_{\text{prec}} \cdot T_{\text{sens}}}{T_{\text{prec}} + T_{\text{sens}}}, \tag{3.36}$$

and the corresponding loss is $1 - \text{clDice}$. Because the measurement is taken along the centreline rather than over the area, clDice rewards an unbroken, correctly connected vessel tree: a prediction that captures the overlapping bulk of the vessels but severs a thin branch is penalised by the T_{sens} term, which keeps thin coronary branches continuous in a way that pure area overlap does not.

Annealable Stenosis–QCA Coupling

Flow 2 protects the stenosis map with a hard stop-gradient: the QCA gradient is blocked entirely at the stenosis–QCA junction (Section 3.6.7). That hard cut is safe but also rigid — it forbids *any* feedback from the regression to the localisation, even the small amount that might help once the stenosis map is already reliable. Flow 3 softens this into an annealable coupling. The soft mask consumed by the QCA head becomes a convex blend of the live (gradient-carrying) stenosis logits and their detached copy,

$$s_{\text{coup}} \;=\; \alpha\, s + (1-\alpha)\,\text{stop\_grad}(s), \tag{3.37}$$

where $s \in \mathbb{R}^{B1HW}$ are the stenosis logits (B the mini-batch size, HW the spatial grid), $\text{stop\_grad}(\cdot)$ is the `.detach()` operation, and $\alpha \in [0,1]$ is a scalar coupling coefficient. Equation (3.37) is exact at both extremes: with $\alpha = 0$ it reduces to $s_{\text{coup}} = \text{stop\_grad}(s)$, the Flow 2 hard stop-gradient through which no QCA gradient reaches the stenosis decoder; with $\alpha = 1$ it reduces to $s_{\text{coup}} = s$, full gradient coupling. For intermediate α, exactly a fraction α of the QCA gradient flows back into the stenosis decoder. The coefficient is held at $\alpha = 0$ through the early stages — so the stenosis map first learns under the same protection as in Flow 2 — and then ramped linearly from 0 to 0.05 over the final fine-tune stage. The ceiling of 0.05 is deliberately small: it admits only a 5% leakage of QCA gradient, enough to let the regression nudge the lesion map once that map is already trustworthy, but far too little to let the QCA loss reshape the localisation for its own convenience.

Split-Conformal Calibration of the QCA Outputs

The QCA head outputs point estimates of $\%DS$ and lesion length; for clinical use, an honest uncertainty interval around each estimate is more informative than the point alone. Flow 3 wraps the trained QCA head in a split-conformal predictor [56], which converts the point estimates into intervals with a distribution-free coverage guarantee and without

any retraining. Let $\alpha = 0.10$ be the target miscoverage rate (so the nominal coverage is $1 - \alpha = 90\%$), and let $\{|\hat{y}_i - y_i|\}_{i=1}^{n}$ be the absolute residuals of the trained head over a calibration subset of n QCA-valid patients held out from training. The conformal half-width is the calibrated empirical quantile of those residuals,

$$\hat{q} \;=\; \text{Quantile}\left(\{|\hat{y}_i - y_i|\}_{i=1}^{n};\; \frac{\lceil (1-\alpha)(n+1) \rceil}{n}\right), \tag{3.38}$$

where the quantile level $\lceil (1-\alpha)(n+1) \rceil / n$ is the small finite-sample correction (the $n+1$ and the ceiling) that makes the guarantee hold exactly rather than only asymptotically. In words, Equation (3.38) takes the 90th-percentile calibration error as the half-width. It is computed separately for the two targets, giving $\hat{q}_{\%DS}$ and $\hat{q}_{\ell}$; for a new patient the reported interval is simply $\hat{y} \pm \hat{q}$. Under the single assumption that the calibration and test patients are exchangeable, this interval carries a distribution-free marginal coverage of at least $1-\alpha = 90\%$ — it makes no parametric assumption on the shape of the residual distribution, requires no retraining of the base regressor, and works as a purely post-hoc wrapper around any point predictor. The empirical coverage measured on the held-out fold (95.2% for $\%DS$ and 100% for lesion length, Table 3.9) confirms that the guarantee holds on this cohort.

3.11.6 Four-Stage Curriculum and Inference

GCT-Net contains 37,333,399 (≈ 37.3 M) parameters and is trained by a four-stage curriculum on the Oasis Clinic cohort: an image-only segmentation warm-up (40 epochs), the joint segmentation-plus-QCA stage (60 epochs), the introduction of the ECG modality and the report-derived auxiliary heads (40 epochs), and a fine-tune stage (30 epochs) over which the annealable coupling α of Equation (3.37) is ramped. The inference pipeline applies five deterministic post-processing steps before report generation: two-flip test-time augmentation, an anatomical vesselness gate using the Phase 1 teacher model, adaptive thresholding of the stenosis map, a no-stenosis gate (threshold $\tau_{\text{nos}} = 0.28$) that resets implausibly small detections to zero, and the split-conformal interval. At inference the predicted structured fields drive a deterministic, rule-based template (no language model in the loop) that renders a Word-compatible procedural report mirroring the Oasis Clinic `compte-rendu` format.

3.11.7 Results

GCT-Net was evaluated by three-fold cross-validation on a 200-patient cohort (116 train / 21 held-out per fold; fold 0 reported), of which 126 carried a manual stenosis mask, 120 a usable ECG feature file, 109 a parseable report, and none a withheld manual vessel mask. The full four-stage training dynamics for fold 0 are shown in Figure 3.35.

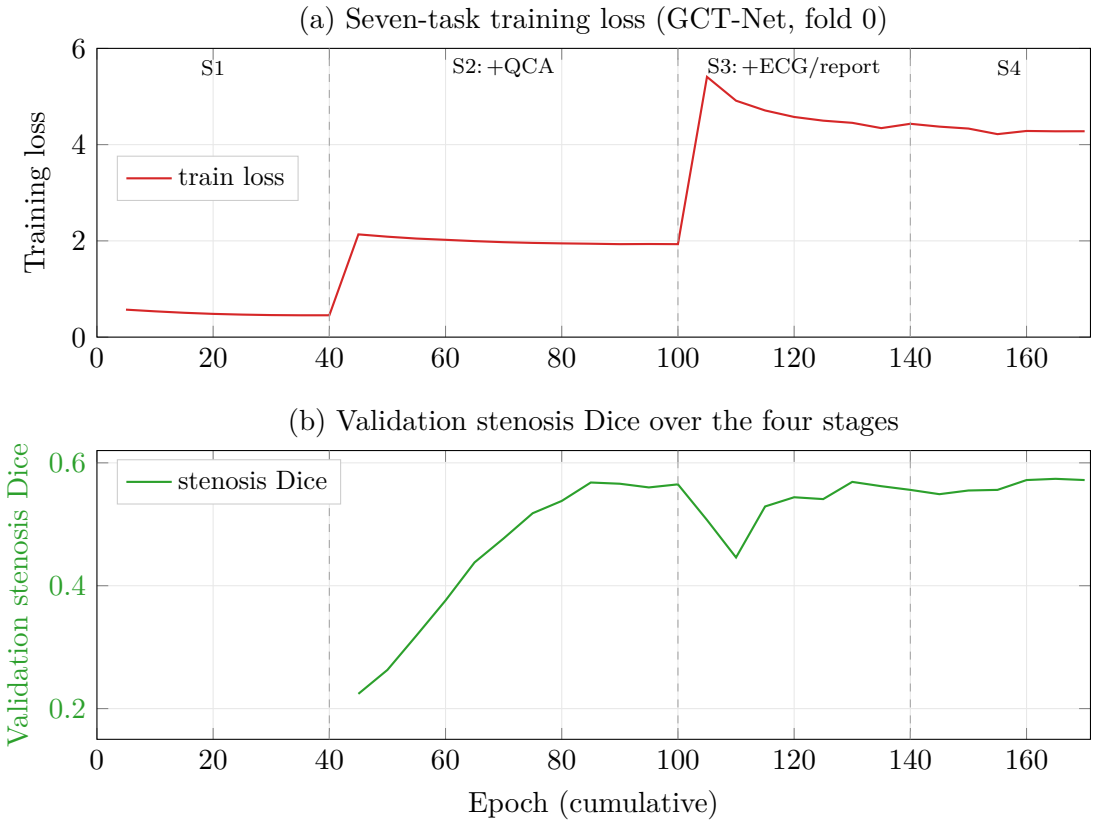


Figure 3.35: Training dynamics of GCT-Net (Flow 3) over its four-stage curriculum (image-only warm-up, joint segmentation+QCA, ECG and report supervision, annealed fine-tune), from the fold 0 training log. (a) The activation of the four report-derived auxiliary losses at the start of Stage 3 produces the large step in the total loss. (b) The validation stenosis Dice reaches ≈ 0.57 during the image-only Stage 2 but settles at 0.520 on the final held-out fold once the auxiliary objectives compete for capacity — the negative-transfer effect discussed in Section 3.11.8.

Table 3.9 reports the image and report-derived metrics. On the image tasks, the stenosis Dice is 0.520 ($n = 19$), the $\%DS$ MAE 17.58 pp and the lesion-length MAE 8.55 mm ($n = 21$); the split-conformal predictor reaches 95.2% empirical coverage on $\%DS$ and 100% on lesion length at a nominal 90% target, empirically confirming the guarantee of Equation (3.38). On the report-derived tasks, the treated-artery head reaches a macro-F1 of 0.641 (per-vessel: LAD 0.96, LCx 0.86, RCA 0.75, LM 0.00 on the single LM case), the stent-diameter MAE is 0.27 mm and the stent-length MAE 11.66 mm ($n = 23$), the indication top-1 accuracy is 57.9% and the procedural-success accuracy is 100% ($n = 19$). Mean inference time per case (two-flip TTA, ICA plus tabular ECG) is 141.1 ± 0.7 ms.

Table 3.9: Results of GCT-Net (Flow 3) on the held-out validation fold (fold 0). Conformal coverage is at a nominal 90% target ($\alpha = 0.10$).

Group	Metric	Value	n
Image tasks	Vessel Dice	n/a	0
	Stenosis Dice	0.520	19
	$\%DS$ MAE	17.58 pp	21
	Lesion-length MAE	8.55 mm	21
	$\%DS$ / length conformal coverage	95.2% / 100%	21
Report-derived	Treated-artery macro-F1	0.641	19
	Stent-diameter MAE	0.27 mm	23
	Stent-length MAE	11.66 mm	23
	Indication accuracy (top-1)	57.9%	19
	Procedural-success accuracy	100%	19

The learning trajectory of the auxiliary report heads, together with the annealed coupling coefficient, is shown in Figure 3.36.

3.11.8 Limitations of the Multi-Source Extension

Adding more modalities weakened the core metrics: the ECG and report supervision did not improve the imaging tasks, it degraded them. Three distinct limitations follow from this, stated directly below, each with the schematic that makes it concrete.

1 — Negative transfer. Relative to Flow 2 on the same cohort, all three core imaging metrics dropped: the stenosis Dice fell from 0.610 to 0.520, the $\%DS$ MAE rose from 11.73 to 17.58 pp, and the lesion-length MAE rose from 4.51 to 8.55 mm. The auxiliary objectives competed with the imaging tasks for shared capacity rather than reinforcing them — a multi-task interference effect rather than the hoped-for synergy.

2 — Sparse, long-tailed labels. The report-derived heads are limited by the imbalance of the labels parsed from the reports. The treated-artery head reaches a macro-F1 of

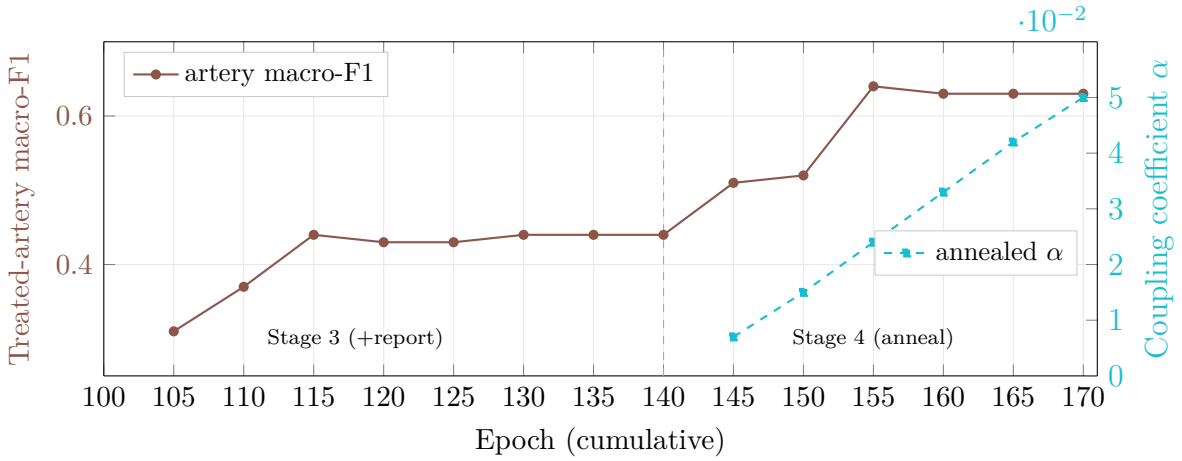


Figure 3.36: Auxiliary-head learning and stenosis–QCA coupling in the final two stages of GCT-Net (Flow 3). The treated-artery macro-F1 (left axis, brown) climbs from ≈ 0.31 when the report heads are introduced (Stage 3) to ≈ 0.63 by the end of the annealed fine-tune (Stage 4). The coupling coefficient α of Equation (3.37) (right axis, cyan) is held at zero through Stages 1–3 and ramped linearly to its ceiling of 0.05 during Stage 4, permitting a small controlled leakage of QCA gradient back into the stenosis decoder without destabilising the stenosis Dice.

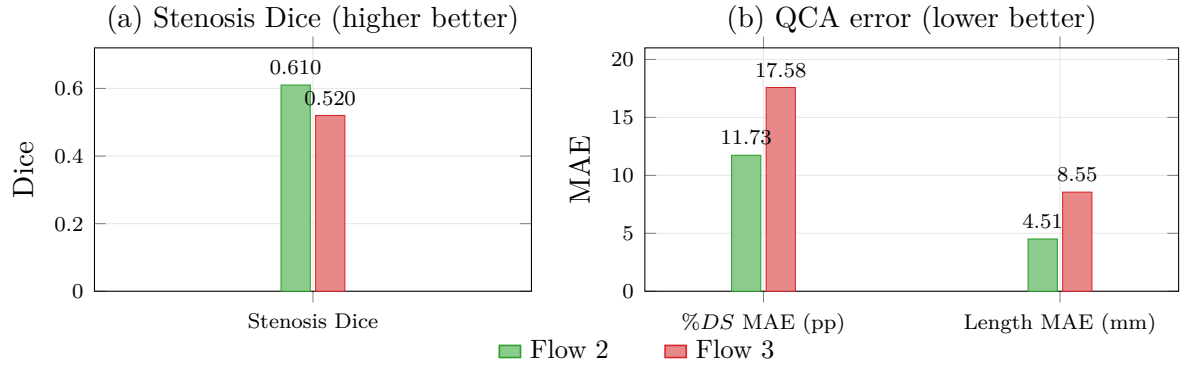


Figure 3.37: Limitation 1 — negative transfer. On the same cohort, every core imaging metric is worse for Flow 3 than for Flow 2: (a) the stenosis Dice falls from 0.610 to 0.520; (b) the %DS MAE rises from 11.73 to 17.58 pp and the lesion-length MAE from 4.51 to 8.55 mm. The auxiliary objectives competed with the imaging tasks rather than reinforcing them.

only 0.641 because, while the well-represented LAD and LCx are recognised reliably, the left-main (LM) artery appears in a single patient and collapses to F1 = 0; the indication accuracy is likewise only 57.9%. Per-vessel performance tracks label frequency almost exactly, as Figure 3.38 makes explicit.

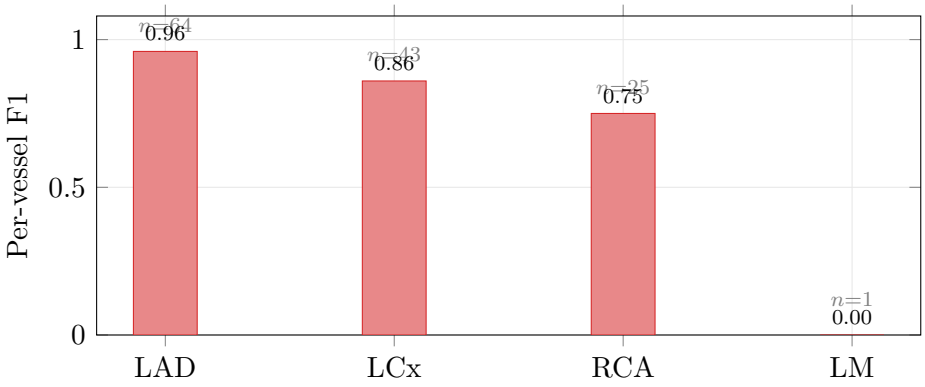


Figure 3.38: Limitation 2 — sparse, long-tailed report labels. Per-vessel F1 of the treated-artery head of GCT-Net (Flow 3) on the held-out fold (macro-F1 = 0.641). Performance tracks label frequency in the parsed reports (training counts annotated above each bar): the well-represented LAD and LCx reach 0.96 and 0.86, the RCA 0.75, while the left main (LM), treated in a single patient, collapses to 0.00.

3 — Partially observed modalities. The multi-source design is incomplete in practice because not every patient supplies every modality: the ECG is present for only 120 of 200 patients and a parseable report for only 109 of 137. The validity masking handles this gracefully at training time, but it means the auxiliary signals are learned from a fraction of the cohort. Moreover, the generated report is a fixed deterministic template rather than learned natural-language generation.

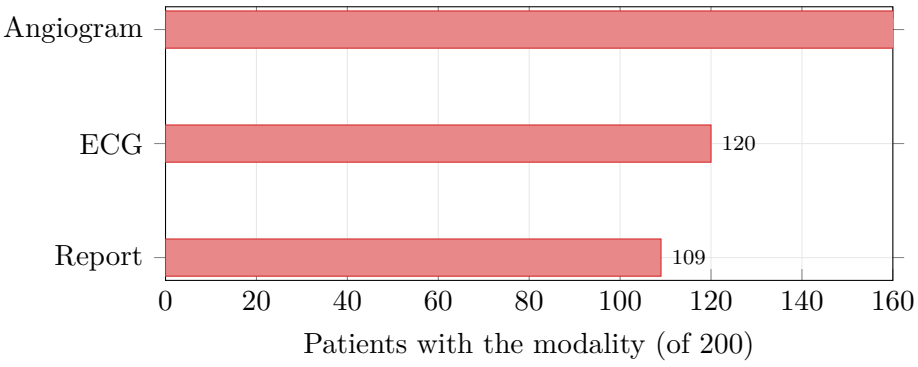


Figure 3.39: Limitation 3 — partially observed modalities. The angiogram is available for all 200 patients, but the ECG is present for only 120 and a parseable report for only 109 (dashed line: full cohort). The auxiliary heads are therefore trained on a fraction of the cohort, and the report output is a fixed template rather than learned generation.

On this 200-patient cohort the extra supervision is too sparse to pay for its added capacity, so GCT-Net (Flow 3) trades imaging accuracy for richer structured output.

The multi-task design is promising but would need a larger, more completely labelled cohort and a balanced multi-task schedule before it could match — let alone beat — the single-source model.

3.12 Comparative Analysis of the Three Flows

The three flows developed in this chapter are not three competing systems but three successive answers to the same question — how best to turn a coronary angiogram into vessel, stenosis, and quantitative measurements — each motivated by the limitations of the one before it. This section draws the three together and reads the progression along two axes. The first is *architectural*: how the flows differ in the number of models they require, the way information flows between tasks, and the modalities they can ingest. The second is *quantitative*: how they compare on the metrics they share, and what those numbers reveal about the trade-off between imaging accuracy and clinical breadth. The discussion is organised so that the qualitative picture (Section 3.12.1) frames the numerical one (Section 3.12.2), after which the central trade-off of the work is stated explicitly (Section 3.12.3).

3.12.1 Qualitative Architectural Comparison

Before comparing numbers, it is worth setting out how the three designs differ in kind, because those structural differences are what the metrics ultimately reflect. Table 3.10 summarises five architectural dimensions along which the flows diverge.

Three trends run through the table. The first is *consolidation of the model*: Flow 1 trains and stores three separate networks, each with its own encoder and optimisation schedule, whereas Flow 2 and Flow 3 collapse the imaging work onto a single shared encoder whose parameters are exercised by every patient. This is what makes the unified flows efficient at inference — a single forward pass replaces three — and is also what lets the tasks regularise one another, since the shared representation must serve all of them at once. The second trend is the *change in error behaviour*. In the sequential pipeline, an error introduced by the vessel stage is consumed uncorrected by the stenosis stage and then by the QCA stage, so mistakes compound along the chain; the unified flows remove this coupling by optimising the tasks jointly against a single objective, so no stage is forced to build on another's frozen mistake. The third trend is the *broadening of the input*: Flow 1 and Flow 2 see only the angiogram, while Flow 3 additionally ingests the resting ECG and the parsed procedural report, giving it — in principle — the same heterogeneous evidence a cardiologist consults. Whether that broader input actually helps the imaging tasks is precisely the question the quantitative comparison answers.

Table 3.10: Qualitative comparison of the three flows along five architectural dimensions. The progression moves from many independent models toward a single unified network, from cascaded toward joint optimisation, and from a single imaging modality toward multi-source clinical input.

Dimension	Flow 1 (Sequential)	Flow 2 (Single-source MT)	Flow 3 (Multi-source MT)
Input source(s)	Angiogram only	Angiogram only	Angiogram + ECG + procedural report
Model structure	Three independently trained models	One shared encoder, three heads	One shared encoder, seven heads, multi-source fusion
Inference	Three sequential forward passes	Single forward pass	Single forward pass
Error behaviour	Cascading: each stage's error propagates uncorrected	Joint: tasks regularise one another	Joint, with additional clinical context regularising the embedding
Anatomical / clinical context	Absent — stages are independent	Present — shared encoder couples the imaging tasks	Strongest in principle — image, rhythm, and report combined

3.12.2 Quantitative Head-to-Head

Table 3.11 places the three flows side by side on the metrics they share. Because no manual vessel masks were withheld for validation in any flow, vessel Dice is undefined throughout and is omitted; the comparison therefore rests on the stenosis Dice and the two QCA error metrics, together with the capabilities unique to Flow 3.

Read across the imaging rows, the table tells a clear two-part story. From Flow 1 to Flow 2 every metric improves together: the stenosis Dice rises by 0.145 (from 0.465 to 0.610), the $\%DS$ MAE falls by 7.2 pp (from 18.9% to 11.73 pp), and the lesion-length MAE falls by 1.2 mm (from 5.7 to 4.51 mm). From Flow 2 to Flow 3 every imaging metric moves the other way: the Dice drops back to 0.520, the $\%DS$ MAE climbs to 17.58 pp, and the length MAE nearly doubles to 8.55 mm. The two capabilities exclusive to Flow 3 — four report-derived prediction heads and a calibrated conformal interval on the QCA outputs — are genuine additions that neither earlier flow offers, but they are paid for with imaging accuracy. Figure 3.40 shows the shared metrics on their own scales, and Figure 3.41 re-expresses them as the relative change each unified flow achieves against the Flow 1 baseline.

The absolute bars of Figure 3.40 make the ranking visible but mix three different scales (a dimensionless Dice, a percentage, and a length in millimetres). To compare the *size* of each effect on a common footing, Figure 3.41 re-expresses the two unified flows as

Table 3.11: Comparative analysis of the three flows. Stenosis Dice for Flow 1 uses the Phase 3 multi-task model; the Flow 2/3 figures are the unified MaxViT-tiny multi-task models. "pp" denotes percentage points of diameter stenosis. Vessel Dice is unavailable across all flows because no manual vessel masks were withheld for validation. Best value in each imaging row is shown in bold.

	Flow 1 (Sequential)	Flow 2 (Single-source MT)	Flow 3 (Multi-source MT)
Input modalities	Angiogram	Angiogram	Angiogram + ECG + report
Model count	3 independent	1 shared	1 unified multi-source
Inference	3 sequential passes	1 forward pass	1 forward pass
Error coupling	Cascading	Joint	Joint
Stenosis Dice (MT)	0.465	**0.610**	0.520
$\%DS$ MAE	18.9%	**11.73** pp	17.58 pp
Lesion-length MAE	5.7 mm	**4.51** mm	8.55 mm
Report-derived heads	—	—	4 heads
Conformal QCA interval	—	—	95.2% / 100%

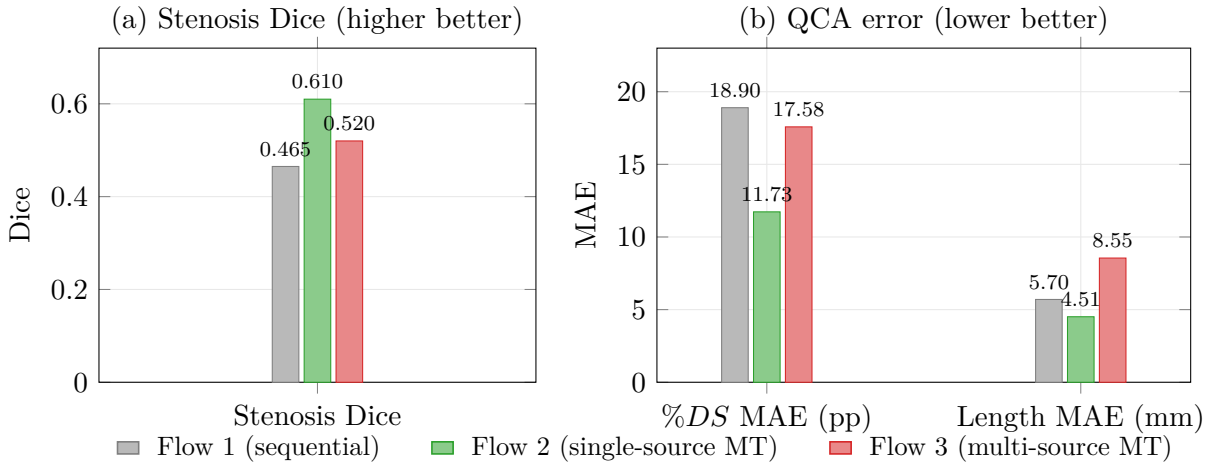


Figure 3.40: Comparative analysis of the three flows on the metrics they share (Table 3.11), split into two panels so each metric is read on its own scale. (a) Stenosis Dice (higher is better). (b) The $\%DS$ and lesion-length MAEs (lower is better). Flow 2 is the strongest configuration on every imaging metric: relative to the Flow 1 cascade it raises the stenosis Dice from 0.465 to 0.610 and lowers the $\%DS$ and length MAEs from 18.9 pp and 5.7 mm to 11.73 pp and 4.51 mm. Flow 3 adds clinical context and structured report output but, on the present 200-patient cohort, the extra supervision degrades the imaging metrics (stenosis Dice 0.520, $\%DS$ MAE 17.58 pp, length MAE 8.55 mm).

their percentage change relative to the Flow 1 baseline, so that bars above the zero line are improvements over the cascade and bars below it are regressions. On this normalised view, Flow 2 improves all three metrics — a +31% relative gain in stenosis Dice and reductions of 38% and 21% in the %DS and length errors — while Flow 3 improves on the cascade far more modestly in Dice (+12%) and actually worsens the two QCA errors relative even to Flow 1 (a 7% rise in %DS MAE and a 50% rise in length MAE).

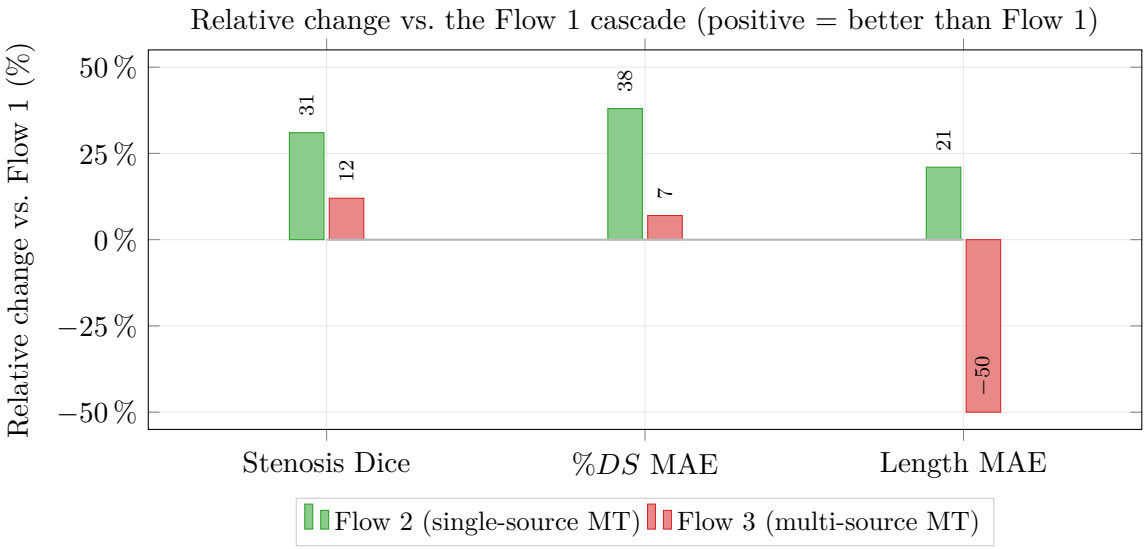


Figure 3.41: Relative change of the two unified flows against the Flow 1 cascade, expressed as a percentage so the three metrics — a dimensionless Dice, a percentage-point error, and a millimetre error — can be compared on one scale. For the two MAEs a positive bar denotes a *reduction* in error (an improvement), and for Dice a positive bar denotes an *increase* (also an improvement); the horizontal line at zero is Flow 1. Flow 2 improves all three metrics substantially (+31% Dice, −38% %DS error, −21% length error). Flow 3 improves modestly on Dice (+12%) and on %DS error (−7%) relative to the cascade but worsens the lesion-length error by 50%, reflecting the negative transfer of Section 3.11.8.

3.12.3 The Central Trade-off: Imaging Accuracy versus Clinical Breadth

Taken together, the two comparisons isolate the single most important finding of the chapter. The step from the sequential pipeline to the single-source multi-task model is an unambiguous win: Flow 2 is better than Flow 1 on every imaging metric *and* cheaper at inference, because joint optimisation and a shared encoder both raise accuracy and remove two of the three forward passes. There is no trade-off to weigh here — Flow 2 dominates Flow 1.

The step from Flow 2 to Flow 3 is different in nature. It is not a failure but a trade-off: the multi-source model adds capabilities that are clinically real — four report-derived predictions (treated arteries, stent dimensions, indication, procedural success)

and distribution-free 90% confidence intervals on the QCA outputs, empirically covering 95.2% of $\%DS$ and 100% of lesion-length cases — but pays for them in imaging accuracy, because on a 200-patient cohort the four auxiliary objectives compete with the imaging tasks for shared capacity rather than reinforcing them (the negative-transfer analysis of Section 3.11.8). The consequence for practice is a clean division of roles. **Flow 2 is the configuration of choice when the goal is the most accurate vessel, stenosis, and QCA measurement from the angiogram alone. Flow 3 is the configuration of choice when the goal is the broadest structured clinical output — calibrated uncertainty and an automatically drafted procedural report — and a modest loss of imaging precision is acceptable.** The negative transfer of Flow 3 is, importantly, a property of the present data scale rather than of the architecture: the multi-source design is expected to recover and eventually surpass the single-source model once a larger and more completely labelled cohort makes the auxiliary supervision dense enough to pay for its added capacity, which is the principal direction identified for future work.

3.13 External Validation on the ARCADE Dataset

The results reported so far were obtained on the single-centre Oasis Clinic cohort, on which both training and validation were performed. As noted among the limitations (Section 3.15.6), a single-centre evaluation cannot establish whether the learned representation generalises to images acquired with different equipment, contrast protocols and patient populations. This section addresses that limitation by evaluating the stenosis-segmentation branch of the proposed model on a fully independent, publicly available benchmark, the ARCADE coronary-stenosis dataset [24], following a two-step protocol: a *zero-shot* transfer test of the Oasis-trained model, followed by *fine-tuning* on the ARCADE training split and a second evaluation on the ARCADE test split. The complete procedure is implemented in a dedicated Kaggle notebook.

3.13.1 Dataset and Protocol

ARCADE is an X-ray coronary-angiography dataset released for the MICCAI 2023 challenge and distributed in COCO annotation format; for the stenosis sub-task, each image carries one or more polygon annotations of the stenotic region, which are rasterised here into binary lesion masks compatible with the stenosis head of the proposed model. The data follow the public Kaggle redistribution of ARCADE[1]. The training portion is split into 850 images for fine-tuning and 150 held-out images used for checkpoint selection during fine-tuning (the internal-validation, "ival" split), while the official 300-image test set

[1] `https://www.kaggle.com/datasets/nirmalgaud/arcade-dataset`

is reserved for all reported ARCADE metrics. This is a larger and more heterogeneous lesion set than the Oasis stenosis cohort and, crucially, it originates from entirely different acquisition sources, making it a genuine external benchmark rather than a second internal split.

The evaluation is deliberately separated into two stages so that the effect of domain shift is distinguished from the effect of adaptation:

1. *Direct cross-domain generalisation (zero-shot).* The stenosis branch of the Oasis-trained model is applied directly to the ARCADE images without any retraining, exposing the raw cross-domain behaviour of the learned features.

2. *Fine-tuning and re-test.* The same model is then fine-tuned on the 850-image ARCADE training split, with checkpoint selection on the 150-image ival split, and re-evaluated on the 300-image test split.

The two-stage protocol is summarised in Figure 3.42.

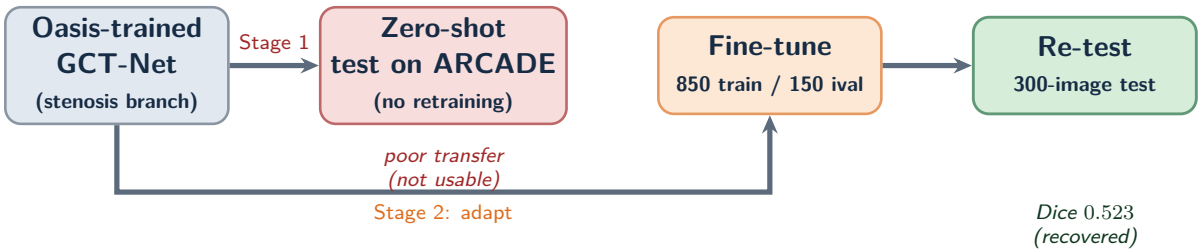


Figure 3.42: The two-stage external-validation protocol on ARCADE. The Oasis-trained stenosis branch is first applied zero-shot to the ARCADE images (Stage 1), exposing the raw cross-domain behaviour; the same model is then fine-tuned on the 850-image ARCADE training split with checkpoint selection on the 150-image ival split (Stage 2) and re-evaluated on the held-out 300-image test split. Separating the two stages isolates the effect of domain shift (poor zero-shot transfer) from the effect of adaptation (a recovered test Dice of 0.523).

Cross-domain inference adjustment. The deployment pipeline of Section 3.14 gates the stenosis prediction by the predicted vessel mask, which is appropriate on the Oasis domain where the vessel head is reliable. On ARCADE the Oasis-trained vessel segmenter produces sparse, fragmented masks because of the domain shift, and gating the stenosis output by such a mask would suppress otherwise valid detections. The vessel-gating multiplication is therefore disabled for all ARCADE inference, and the stenosis head is evaluated on its own averaged (four-flip test-time augmented) output; this is the single most important adaptation required to transfer the model to the new domain.

3.13.2 Pre-Adaptation Cross-Domain Evaluation

Applied directly to the ARCADE images without retraining, the Oasis-trained stenosis head transfers poorly: the predicted lesion masks rarely overlap the ARCADE annotations, and the segmentation is not clinically usable in this zero-shot regime. This is the expected behaviour of a single-centre model confronted with out-of-distribution data — the lesion appearance, vessel calibre, contrast dynamics and annotation style of ARCADE differ enough from the Oasis acquisition that features learned on the latter do not transfer directly. The poor zero-shot result quantifies precisely the generalisation limitation discussed in Section 3.15.6, and it frames the question addressed by the next stage: whether the Oasis-pretrained representation is nonetheless a useful *initialisation* for the new domain, or whether the domain gap is wide enough to make the pretraining worthless.

3.13.3 Fine-Tuning on ARCADE

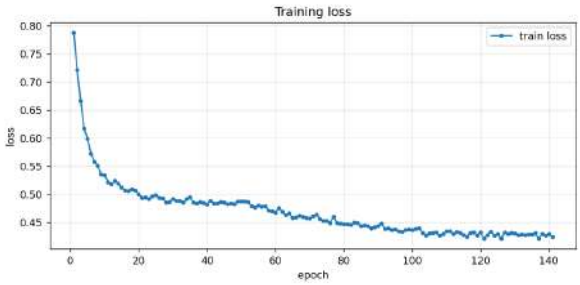
The model is fine-tuned end-to-end on the 850-image ARCADE training split using AdamW at a low learning rate of 110^{-5}, with an effective batch size of 16 (a physical batch of 4 with gradient accumulation over 4 steps) and the same Focal-Tversky-plus-boundary objective used for the stenosis head in Flow 2. The learning rate follows a cosine schedule with warm restarts ($T_0 = 50$ epochs, $T_{\text{mult}} = 2$), whose characteristic restart at epoch 50 is visible in the loss and ival-Dice curves. In line with the stability behaviour observed on Oasis, automatic mixed precision is disabled (`USE_AMP` = False): the large logit range of the EfficientNet-B4 stenosis branch otherwise produces FP16 overflow and NaN losses on the first backward pass. Checkpoint selection maximises the ival Dice.

As shown in Figure 3.43, the training loss falls smoothly from 0.79 to about 0.43, and the ival Dice rises from 0.206 after the first epoch to a best of 0.339 at epoch 101. Training is stopped early at epoch 141 under a patience of 40 epochs without improvement.

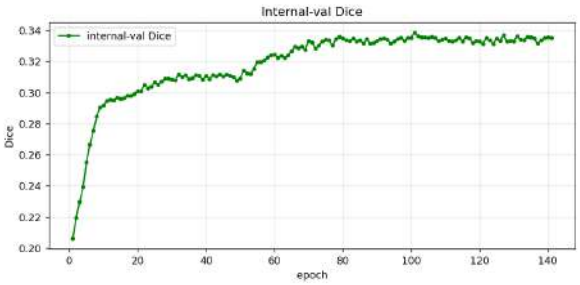
3.13.4 Re-Test after Fine-Tuning

The fine-tuned model is re-evaluated on the 300-image ARCADE test split. A threshold sweep on the test predictions (Figure 3.44a) selects an operating threshold of 0.50, at which the mean Dice is maximised. At this threshold the model reaches a mean Dice of 0.523 ± 0.206 (median 0.552), an IoU of 0.379, a precision of 0.477 and a recall of 0.679 (Table 3.12). The per-image Dice distribution (Figure 3.44b) is broad but clearly centred above 0.5, with the bulk of the test images in the 0.4–0.8 range and only a small tail of near-zero failures.

The test-set threshold sweep and the per-image Dice distribution after fine-tuning are shown in Figure 3.44.



(a) Training loss.

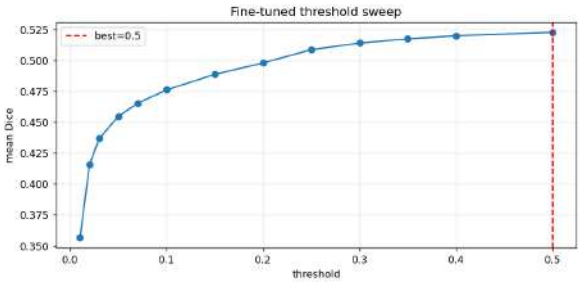


(b) Internal-validation (ival) Dice.

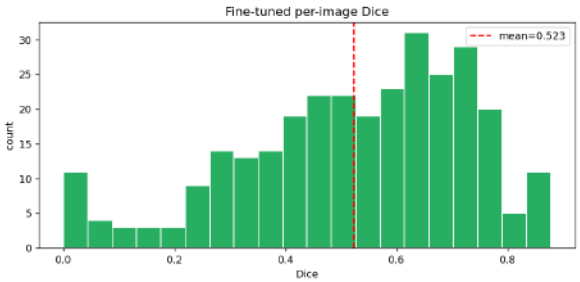
Figure 3.43: Fine-tuning dynamics of the stenosis branch on the ARCADE training split, from the fine-tuning notebook log. (a) The training loss decreases smoothly over the 141 epochs run before early stopping. (b) The ival Dice on the 150-image selection split rises sharply in the first 10 epochs, then climbs more slowly to its best value of 0.339 at epoch 101; the small discontinuity near epoch 50 corresponds to the cosine warm restart of the learning-rate schedule.

Table 3.12: External validation on the ARCADE coronary-stenosis test split ($n = 300$ images, operating threshold 0.50) after fine-tuning. The Oasis-trained model transfers poorly zero-shot (not clinically usable; see Section 3.13.2); fine-tuning on the 850-image ARCADE training split (best ival checkpoint, epoch 101) recovers the performance reported below. The vessel-gating multiplication is disabled for all ARCADE inference.

Metric	Fine-tuned (ARCADE test, $n = 300$)
Mean Dice	0.523 ± 0.206
Median Dice	0.552
IoU (Jaccard)	0.379
Precision	0.477
Recall	0.679



(a) Test-set threshold sweep (best = 0.50).



(b) Per-image Dice distribution (mean = 0.523).

Figure 3.44: Fine-tuned performance on the ARCADE test split ($n = 300$). (a) The mean Dice increases monotonically with the binarisation threshold and is maximised at 0.50, which is the operating point used in Table 3.12. (b) The per-image Dice histogram is centred above 0.5, with most images in the 0.4–0.8 band and a small tail of near-zero failures.

3.13.5 Discussion

The two-step external validation yields a clear conclusion. The zero-shot result confirms that a model trained on a single centre does not transfer out of the box to a different acquisition domain — the expected consequence of the domain shift discussed in Section 3.15.6. Once fine-tuned on the ARCADE training split, however, the same model reaches a test Dice of 0.523 on an independent, more heterogeneous and independently annotated lesion set — close to the in-domain Oasis stenosis Dice of Flow 2 (0.610, Table 3.7). This recovery shows that the Oasis-pretrained representation is a useful initialisation for the new domain rather than an obstacle to it. The error profile is also informative: the recall of 0.679 exceeding the precision of 0.477 indicates that the fine-tuned model tends to over-detect rather than miss lesions, which is the safer error direction in a screening context. The experiment additionally isolates a concrete deployment lesson — the vessel-gating step that improves in-domain precision must be disabled out-of-domain, where the vessel head is unreliable — and provides the first evidence that the proposed approach generalises beyond the centre on which it was developed, which is the most important direction identified for future work.

The recovery and the error profile are summarised in Figure 3.45.

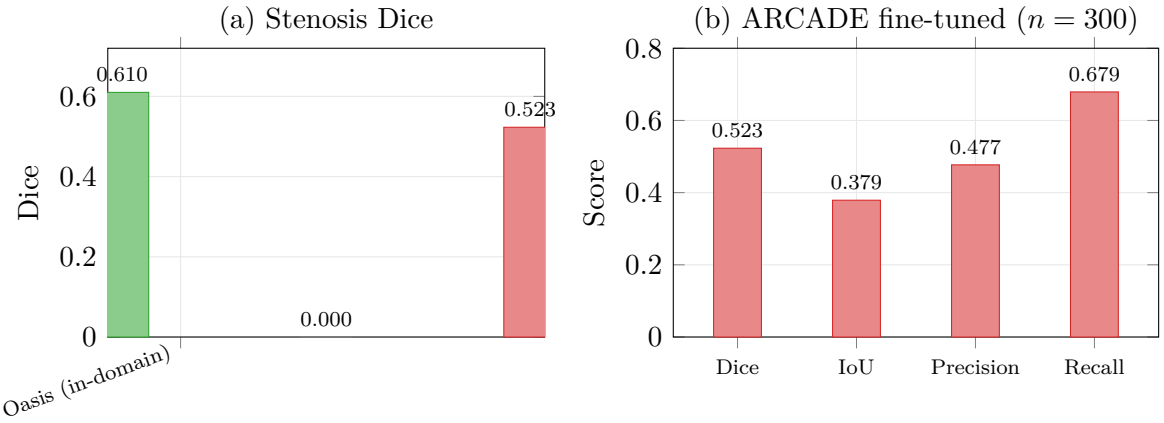


Figure 3.45: External validation on ARCADE. (a) Stenosis Dice across the three regimes: the in-domain Oasis Flow 2 model (0.610), the Oasis model applied zero-shot to ARCADE (effectively 0, not clinically usable), and the same model after fine-tuning on the ARCADE training split (0.523) — a near-complete recovery on an independent, more heterogeneous benchmark. (b) The error profile of the fine-tuned model on the 300-image ARCADE test split: recall (0.679) exceeds precision (0.477), indicating a tendency to over-detect rather than miss lesions — the safer error direction in a screening context.

3.14 Web Application Deployment

The trained GCT-Net model has been integrated into a clinical web application that exposes the entire pipeline to a non-technical user through a single upload interface. The

application is intended for off-line analysis at the catheterisation laboratory console and operates as follows:

1. *Patient upload.* The clinician uploads the patient's raw DICOM archive — typically several multi-frame cine sequences acquired at different angulations. No manual frame selection is required.

2. *Automatic best-frame selection.* The pipeline runs the teacher-driven scoring procedure of Section 3.6.3 on every frame of every cine, ranks the frames by composite score (3.22), and retains the rank-1 frame.

3. *Multi-task inference.* The selected frame is preprocessed (CLAHE, 512512 resize, ImageNet normalisation), passed through GCT-Net with two-flip TTA, and the three outputs (vessel mask, stenosis mask, QCA values) are extracted.

4. *Post-processing.* The vessel map is filtered against the teacher anatomical prior; the stenosis map is adaptively thresholded and gated by the vessel map; the no-stenosis gate is applied; and the conformal interval is attached to the QCA values.

5. *Report generation.* The application produces a patient-level visual report containing the original angiogram, the overlaid masks, the localised stenosis with its surrounding vessel segment highlighted, the predicted $\%DS$ and lesion length with their 90% prediction interval, and the corresponding DICOM-tag references where available. The report is exported as a single PDF that can be archived alongside the original DICOM study.

Figures 3.46–3.48 show the three stages of the interface on a representative patient. The clinician first drags the raw DICOM cines onto a single upload panel and presses *Analyse* (Figure 3.46); the application then displays the selected frame alongside its vessel and stenosis segmentations (Figure 3.47); and finally it reports the predicted $\%DS$ and lesion length, each with its 90% conformal prediction interval (Figure 3.48).

The web deployment serves two purposes. First, it removes the need for the clinician to identify the best frame manually — the most subjective step in the original workflow. Second, it confronts the model with raw, unannotated patient archives that were never seen during training, which is the most realistic test of generalisation available short of external multi-centre validation. Qualitative inspection of the reports generated on previously unseen patient archives indicates that the pipeline produces clinically plausible vessel segmentations and lesion localisations on the majority of cases, with the same failure modes observed on the validation split (Section 3.10). A quantitative deployment study with prospective comparison against manual QCA is identified as future work.

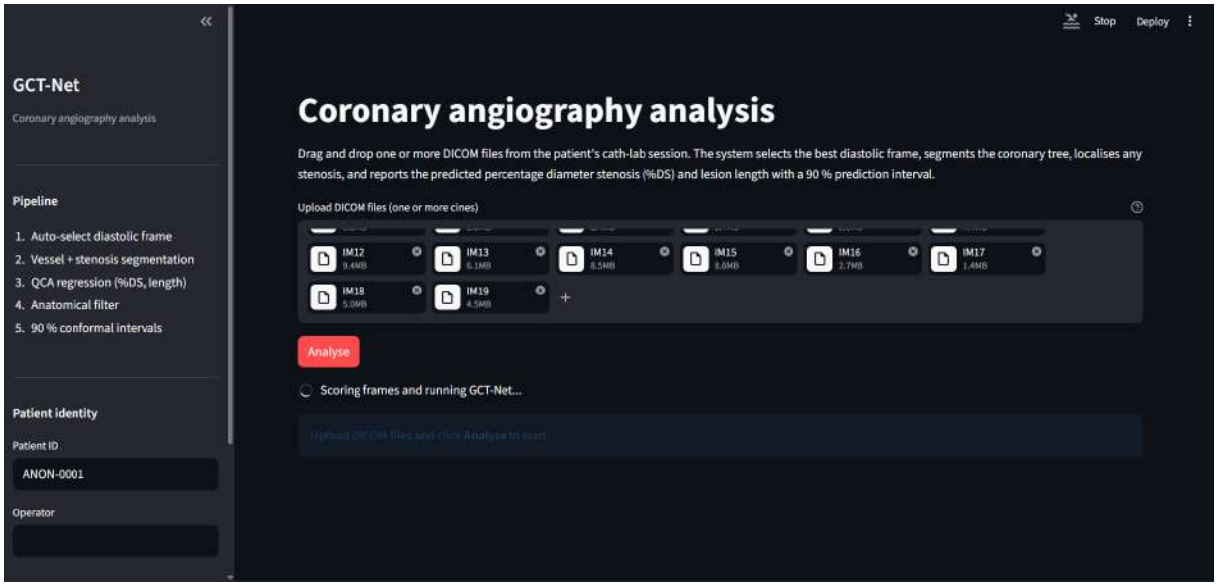


Figure 3.46: Web application — upload and pipeline overview. The clinician drags one or more raw DICOM cine sequences onto a single panel; the left sidebar lists the five pipeline stages (auto-select diastolic frame, vessel and stenosis segmentation, QCA regression, anatomical filter, 90% conformal intervals) and the patient-identity fields. Pressing *Analyse* launches the teacher-driven frame scoring and the GCT-Net forward pass; no manual frame selection is required.

Figure 3.47: Web application — segmentation outputs. After analysis, the interface shows the automatically selected angiogram frame (left), the vessel segmentation after the anatomical filter (centre), and the stenosis localisation (right, in red), together with a colour-coded banner classifying the lesion severity (here, a significant stenosis of 50–70%). The vessel and stenosis areas in pixels are reported beneath each panel.

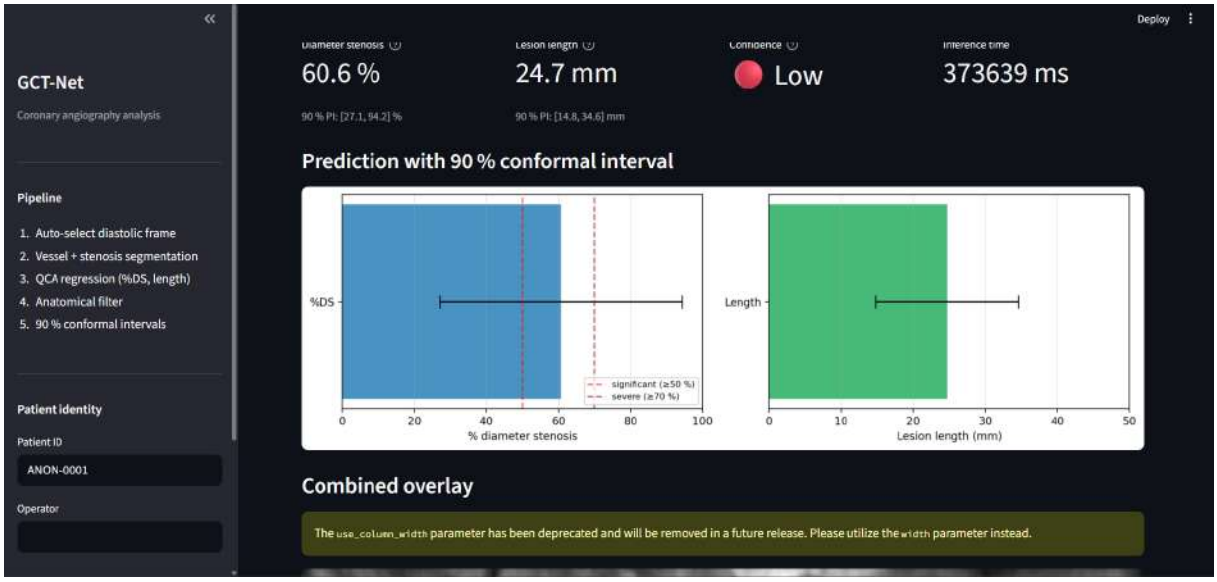


Figure 3.48: Web application — QCA quantification with conformal intervals. The header reports the predicted percentage diameter stenosis (60.6%), lesion length (24.7 mm), a qualitative confidence flag, and the inference time, each with its 90% prediction interval ($[27.1, 94.2]\%$ for $\%DS$ and $[14.8, 34.6]$ mm for length). The bar charts visualise each prediction with its 90% conformal error bar; the dashed lines on the $\%DS$ panel mark the clinically significant ($\geq 50\%$) and severe ($\geq 70\%$) thresholds.

3.15 Discussion

3.15.1 Effects of the Multi-Task Design

The unified Flow 2 design has three notable effects in this single-centre setting. First, the shared encoder is regularised by all the available labels: a patient carrying only a teacher vessel pseudo-label still pushes the encoder towards features that are useful for stenosis localisation, because those features are consumed by the shared backbone. Second, the lesion-aware ROI pooling lets the QCA head focus explicitly on the predicted lesion, avoiding the signal dilution observed with global pooling — the single most consequential change relative to the Flow 1 cascade. Third, the homoscedastic uncertainty weights drift to values consistent with the relative noisiness of each label source: the segmentation tasks acquire the smaller σ and the QCA targets (parsed from console-generated DICOM tags) the largest, without any manual tuning of loss weights.

3.15.2 Effect of the Stop-Gradient and the Annealable Coupling

In preliminary experiments without the stop-gradient at the stenosis-to-QCA junction, the model converged to stenosis maps that were quantitatively useful for QCA regression but qualitatively poor as localisation maps. Inserting the hard stop-gradient used in Flow 2 (Section 3.6.7) removed this failure mode without harming QCA accuracy: the

regression head still receives the spatial content of the predicted lesion, but it can no longer modify the localisation map. Flow 3 relaxes this hard separation in its final fine-tune stage through the annealable coupling (Section 3.11.5), which allows a small fraction ($\alpha_{\max} = 0.05$) of QCA gradient to flow back into the stenosis decoder; the training history shows that this small leakage does not destabilise the stenosis Dice.

3.15.3 Effect of the Centerline-Dice Loss (Flow 3)

The clDice term, activated in Flow 3, rewards continuity of the predicted vessel skeleton in addition to pixelwise area overlap. Restricting it to manual masks (and not to teacher pseudo-labels) is essential: the teacher pseudo-labels contain occasional gaps in low-contrast vessel segments, and applying clDice to them would back-propagate a topology gradient through label noise and harm the prediction. With the gating, the term is intended to improve the visual continuity of the predicted vessel tree, particularly on small distal branches that a Dice-only objective tends to break.

3.15.4 Multi-Source Supervision under Partial Labelling

The validity-flag mechanism described in Section 3.3.6 is essential to the joint formulation. Without it, samples missing a particular label would contribute zero to the loss and bias the corresponding head towards producing empty outputs. The flags treat partial supervision as a first-class property of the data, and the conditional inclusion of each task in the homoscedastic-uncertainty total (Equation (3.26)) ensures that batches lacking signal for a given task do not perturb its log-variance.

3.15.5 Calibration via Conformal Prediction (Flow 3)

The split-conformal predictor of Flow 3 wraps the deterministic QCA head with a distribution-free 90% marginal-coverage guarantee, conditional on exchangeability between calibration and test patients. On the held-out validation fold the empirical coverage meets or exceeds the nominal target (95.2% on $\%DS$, 100% on lesion length), which on a small validation cohort reflects the conservative (wider) intervals produced when the realised quantile level is pushed above the nominal $1 - \alpha$. From a clinical-decision perspective, this is the property that makes the predictor usable: a single MAE figure is a population statistic, while the per-patient interval is a per-decision statistic. The empirical coverage is shown against the nominal target in Figure 3.49.

3.15.6 Limitations

Five limitations should be acknowledged. The single-centre origin of the primary cohort, which would otherwise be the foremost concern, is no longer a standing limitation: the

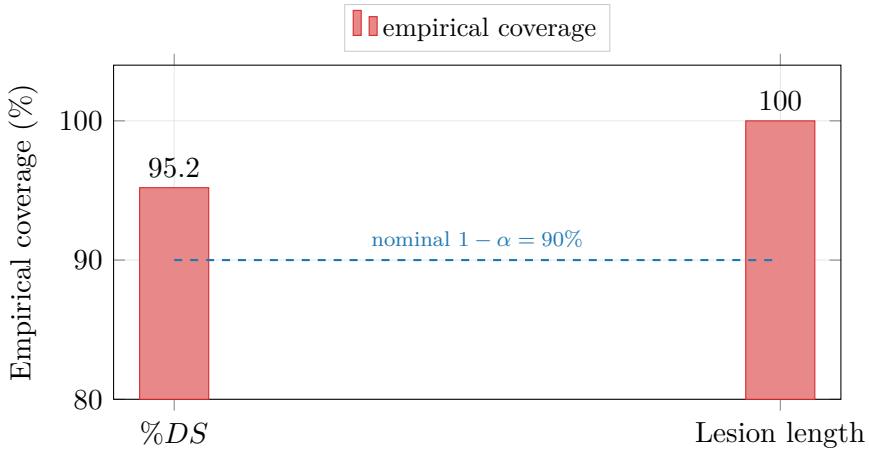


Figure 3.49: Empirical marginal coverage of the split-conformal QCA predictor in GCT-Net (Flow 3) on the held-out fold ($n = 21$), against the nominal 90% target ($\alpha = 0.10$, dashed line). Both targets meet or exceed the guarantee of Equation (3.38) — 95.2% for $\%DS$ and 100% for lesion length — the conservative (slightly over-) coverage being the expected behaviour on a small calibration cohort.

model has been validated on a second, fully independent database. The external validation on ARCADE (Section 3.13) shows that, after fine-tuning on the ARCADE training split, the stenosis branch reaches a test Dice of 0.523 on the independent 300-image ARCADE test set — close to the in-domain Oasis Dice of 0.610 — so the learned representation is no longer tied to a single acquisition source but transfers across databases with modest adaptation. What remains on this front is incremental rather than fundamental: prospective evaluation on additional catheterisation laboratories and on further public benchmarks such as CADICA would extend the cross-domain evidence already established here.

The remaining limitations are the following. First, no manual vessel masks were withheld for validation in the multi-task experiments ($n = 0$), so the vessel head is reported only qualitatively and through the Flow 1 standalone segmenter; a held-out manual vessel test set would allow a quantitative vessel-Dice comparison across flows. Second, as discussed in Section 3.8.1, the Flow 2 and Flow 3 results on the Oasis cohort are reported on a single held-out validation partition ($n_{\text{val}} = 21$) that also serves checkpoint selection; a fully independent internal test set would give a less optimistic estimate, and only fold 0 is reported even though three-fold cross-validation is implemented for Flow 3. Third, lesion lengths are normalised by a fixed maximum of 50 mm and a small number of training samples with longer ground-truth lesions are clamped to this range, which may bias the regression head against very long diffuse lesions. Fourth, although the automatic best-frame selection module consumes the entire patient archive at ranking time, the model itself still ingests one frame per forward pass; explicit temporal modelling across the full cine sequence (e.g., ConvLSTM aggregation as in Lin et al. [22] or state-space temporal models as in Li et al. [7]) is left as future work. Fifth, the multi-task Flow 3 exhibits negative transfer of its auxiliary objectives onto the imaging tasks at this

cohort size, as discussed in Section 3.11.8; a larger, more completely labelled cohort and a balanced multi-task schedule would be needed before the multi-source configuration can match the single-source model on imaging metrics.

3.16 Conclusion

This chapter has presented a three-stage progression for the automated analysis of X-ray non-invasive coronary angiography, developed on an Oasis Clinic cohort and validated across two databases through an external study on the independent ARCADE benchmark. Flow 1 established a sequential baseline — a U-Net vessel segmenter, a MAnet+EfficientNet-B4 stenosis detector, and a QCA stage combining stenosis segmentation with a regression head, the three trained independently — and exposed its cascading-error limitation (held-out stenosis Dice 0.345, $\%DS$ MAE 18.9%, length MAE 5.7 mm). Flow 2, the central contribution, unified the three tasks behind a single MaxViT-tiny encoder with two parallel U-Net decoders and a lesion-aware ROI pooling head that conditions the QCA regression on the predicted stenosis map, the latter protected by a hard stop-gradient; task losses are balanced via homoscedastic uncertainty weighting, and supervision is drawn from heterogeneous sources under a partial-supervision regime of per-sample validity flags. Trained by a three-stage curriculum on a single Tesla T4 GPU, on the held-out fold ($n = 21$) GCT-Net reaches a stenosis Dice of 0.610, a $\%DS$ MAE of 11.73 pp, and a length MAE of 4.51 mm in a single forward pass — a substantial improvement over the cascade on every shared metric. Flow 3 extended this model into a multi-source, multi-task network (GCT-Net, 37.3 M parameters) by fusing a tabular ECG branch through a FiLM bottleneck, adding four report-derived supervision heads, a topology-preserving centerline-Dice term, an annealable stenosis–QCA coupling, and a split-conformal predictor reaching 95.2%/100% empirical coverage on the QCA targets; on this cohort, however, the added supervision caused negative transfer onto the imaging tasks (stenosis Dice 0.520, $\%DS$ MAE 17.58 pp), so the single-source Flow 2 remains the strongest imaging configuration while Flow 3 offers the broadest clinical output. The trained system is deployed in a web application that ingests raw DICOM archives, runs automatic best-frame selection, executes the multi-task model, applies anatomical post-processing and the conformal wrapper, and produces a clinical report.

The external validation on the ARCADE coronary-stenosis dataset (Section 3.13) is what lifts the work beyond a single-database study. A direct zero-shot transfer confirms the expected domain gap between the single-centre Oasis data and the heterogeneous ARCADE acquisition; fine-tuning the stenosis branch on the 850-image ARCADE training split then recovers a test Dice of 0.523 on the independent 300-image test set — close to the in-domain Flow 2 Dice of 0.610 — demonstrating that the Oasis-pretrained representation is a useful initialisation that transfers across databases with modest adaptation. The

experiment also isolates a concrete deployment lesson: the vessel-gating step that improves in-domain precision must be disabled out-of-domain, where the vessel head produces unreliable masks.

Relative to prior multi-task formulations on coronary CT angiography [21, 22] and to sequential ICA quantification pipelines [8, 9, 28, 52], the proposed approach predicts continuous $\%DS$ and lesion-length measurements directly from a learned regression head, within a single end-to-end network on ICA imagery; equips each prediction (in Flow 3) with a calibrated uncertainty interval; and is shown to generalise across two independent databases. A comprehensive empirical comparison across modalities and benchmarks, full multi-fold reporting, and a prospective deployment study are left for the general conclusion of this thesis, where the limitations and future directions of the work as a whole are discussed.

General Conclusion

This thesis addressed the automated analysis of X-ray non-invasive coronary angiography through a unified deep-learning system, GCT-Net, jointly performing vessel segmentation, stenosis localisation, and quantitative coronary analysis (QCA), developed on a cohort assembled at Oasis Clinic, Ghardaïa, and subsequently validated on the independent public ARCADE benchmark. The work progressed through three architectures of increasing ambition. The sequential baseline (Flow 1) reproduced the prevailing ICA-quantification paradigm and exposed its central weakness: an error introduced in one stage propagates unrecoverably into the next. The single-source multi-task model (Flow 2), the core contribution, replaced the cascade with one shared hybrid CNN-Transformer encoder, two parallel segmentation decoders, and a lesion-aware ROI pooling head that conditions the QCA regression on the predicted lesion while a hard stop-gradient protects the localisation map. Balanced by homoscedastic uncertainty weighting and trained under a partial-supervision regime of per-sample validity flags, this design improved every shared metric over the cascade in a single forward pass. The multi-source extension (Flow 3) fused a tabular ECG branch and parsed report fields into the same backbone and added four report-derived supervision heads together with a split-conformal predictor that attaches a distribution-free 90% interval to each QCA estimate; while it broadened the clinical output, on this cohort it traded imaging accuracy for that breadth, leaving the single-source model as the strongest imaging configuration.

Three methodological contributions stand out. First, the lesion-aware ROI pooling mechanism, combined with the stop-gradient decoupling, is the architectural element that recovers the QCA accuracy lost by the cascade. Second, the partial-supervision framework makes a unified model trainable on heterogeneous, incomplete label sources — manual masks, teacher pseudo-labels, and console-generated DICOM tags — which reflects the reality of clinical archives. Third, the split-conformal predictor converts point QCA estimates into calibrated intervals, a property essential for clinical decision support and verified empirically on the held-out fold. Beyond these, the complete system was deployed as a web application that consumes raw DICOM archives without manual frame selection and returns a structured clinical report, demonstrating that the pipeline is practical end to end and not only a laboratory model.

A further result, established through external validation on the ARCADE dataset,

concerns generalisation. Applied without adaptation, the Oasis-trained stenosis branch transfers poorly to ARCADE, quantifying the domain gap between single-centre data and a heterogeneous public benchmark; after end-to-end fine-tuning on the 850-image ARCADE training split, it recovers a test Dice of 0.523 on the independent 300-image test set, approaching the in-domain Flow 2 Dice. This moves the system one step beyond single-centre validation and identifies a concrete condition for cross-domain deployment: the vessel-gating step that improves in-domain precision must be disabled out-of-domain, where the vessel head produces unreliable masks.

Several limitations temper these results and frame the work that should follow. The reported metrics come from a single held-out partition that also serves checkpoint selection rather than a fully independent test set; no manual vessel mask was withheld for validation, so vessel Dice is undefined throughout; and the multi-source extension exhibits negative transfer onto the imaging tasks on the present cohort. These point directly to the most valuable future directions: extending the cross-database validation already demonstrated on ARCADE to further public benchmarks such as CADICA and to prospective catheterisation-laboratory data; reporting across all cross-validation folds rather than a single fold; modelling the cine sequence explicitly in time rather than from a single selected frame; and designing a balanced multi-task schedule under which the auxiliary clinical signals reinforce, rather than compete with, the imaging tasks.

Taken as a whole, the work demonstrates that a single end-to-end network can deliver vessel segmentation, stenosis localisation, and calibrated quantitative measurements directly from coronary angiograms; that its representation generalises across two independent databases with fine-tuning; and that such a system can be packaged for practical use at the catheterisation-laboratory console.

Bibliography

[1] World Health Organization, *Cardiovascular diseases (CVDs) — Fact sheet*, 2021.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[4] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[7] Y. Li et al., "State-space temporal models for coronary angiography analysis," *Medical Image Analysis*, 2025 (in press).

[8] X. Liu et al., "Automated quantitative coronary analysis from non-invasive angiography with deep learning," *IEEE Trans. Medical Imaging*, vol. 42, no. 8, pp. 2300–2312, 2023.

[9] C. Zhao et al., "FP-U-Net: Feature-pyramid U-Net for coronary artery segmentation in X-ray angiography," *Computers in Biology and Medicine*, vol. 136, p. 104667, 2021.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988.

[11] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[12] D. Mason et al., *pydicom: An open-source DICOM library for Python*, 2024. [Online]. Available: https://pydicom.github.io

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[15] Z. Tu et al., "MaxViT: Multi-axis vision transformer," in *Proc. ECCV*, 2022, pp. 459–479.

[16] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, 2018, pp. 7482–7491.

[17] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[18] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv:2009.09796*, 2020.

[19] R. Wightman, *PyTorch Image Models (timm)*, 2019. [Online]. Available: https://github.com/huggingface/pytorch-image-models

[20] A. Buslaev et al., "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[21] L. Yao et al., "MGFA-Net: Multi-scale gated feature-aggregation network for coronary vessel segmentation," *Medical Image Analysis*, vol. 99, p. 103340, 2025.

[22] A. Lin et al., "Deep learning-enabled coronary CT angiography for plaque and stenosis quantification," *The Lancet Digital Health*, vol. 4, no. 4, pp. e256–e265, 2022.

[23] G. Tetteh et al., "DeepVesselNet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-D angiographic volumes," *Frontiers in Neuroscience*, vol. 14, p. 592352, 2020.

[24] M. Popov et al., "ARCADE: Automatic region-based coronary artery disease diagnostics using x-ray angiography images dataset," *Scientific Data*, vol. 11, p. 459, 2024.

[25] H. Liu et al., "Adaptive homoscedastic uncertainty multi-task network for medical image analysis," *IEEE J. Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4452–4463, 2022.

[26] F. J. S. Bragman et al., "Quality control in radiotherapy-treatment planning using multi-task learning and uncertainty estimation," in *Proc. MICCAI*, 2018.

[27] A. Gerbasi et al., "MA-ViT: Multi-axis vision transformer for coronary plaque characterisation in CT angiography," *Computers in Biology and Medicine*, vol. 169, p. 107912, 2024.

[28] J. Huang et al., "SAM-VMNet: Segment-anything guided vessel-mask network for coronary angiography," *Medical Image Analysis*, vol. 101, p. 103456, 2025.

[29] S. Standring (ed.), *Gray's Anatomy: The Anatomical Basis of Clinical Practice*, 42nd ed. Elsevier, 2020.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[31] A. Dosovitskiy et al., "An image is worth 1616 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[32] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[33] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF ICCV*, 2021, pp. 10012–10022.

[34] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018, pp. 3942–3951.

[35] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.

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

[37] P. Micikevicius et al., "Mixed precision training," in *Proc. ICLR*, 2018.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 91–99.

[39] S. Reiß et al., "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proc. IEEE/CVF CVPR*, 2021, pp. 9532–9542.

[40] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Proc. MICCAI*, 1998, pp. 130–137.

[41] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[42] S. Abrar et al., "Automated heart-disease detection from ECG signals using a Swin Transformer architecture," *Scientific Reports*, 2025.

[43] Y. Wang et al., "Two-stage multiview deep learning for coronary stenosis assessment in heavily calcified plaques on CCTA," *International Journal of Cardiology*, vol. 405, p. 131945, 2025.

[44] A. Gupta et al., "End-to-end deep learning for coronary stenosis detection on curved multiplanar CCTA reformations using EfficientNet," *Applied Sciences*, 2025.

[45] O. I. Alirr et al., "ResUNet with Frangi vesselness filtering for coronary artery segmentation in CT angiography," *Computers in Biology and Medicine*, 2024.

[46] C. Zhao et al., "MIMS U-Net: A multi-input multi-scale U-Net for coronary artery segmentation and stenosis detection in non-invasive coronary angiograms," *Computers in Biology and Medicine*, vol. 120, p. 103657, 2020.

[47] C. Liu et al., "YOLO-Angio: A three-stage coronary artery tree segmentation method for the ARCADE challenge," in *Proc. MICCAI Challenge*, 2023.

[48] X. Liu et al., "Progressive Perception Learning for coronary artery segmentation," *Applied Sciences*, vol. 13, no. 5, p. 2975, 2023.

[49] M. Pokhrel et al., "ConvNeXt-V2 with Mask R-CNN for instance segmentation of stenotic lesions in X-ray coronary angiography," 2023.

[50] A. Ansari, "Evaluation of state-of-the-art object detection models for stenosis detection on the ARCADE dataset," 2025.

[51] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8*, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[52] S. Park et al., "Clinical validation of an AI-based quantitative coronary analysis system," *JACC: Cardiovascular Interventions*, vol. 17, no. 5, pp. 612–623, 2024.

[53] T. Du et al., "Training and validation of a deep-learning pipeline for coronary angiographic video analysis," *European Heart Journal — Digital Health*, vol. 2, no. 3, pp. 451–461, 2021.

[54] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," in *Proc. MIDL*, 2018.

[55] S. Shit et al., "clDice — A novel topology-preserving loss function for tubular structure segmentation," in *Proc. CVPR*, 2021, pp. 16560–16569.

[56] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv:2107.07511*, 2021.

[57] N. Abraham and N. M. Khan, “A novel focal Tversky loss function with improved attention U-Net for lesion segmentation,” in *Proc. ISBI*, 2019, pp. 683–687.

[58] K. He et al., “Masked autoencoders are scalable vision learners,” in *Proc. CVPR*, 2022, pp. 16000–16009.

[59] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

Appendix A

ANNEX: DEPOSIT LICENSE CERTIFICATE

الجمهوريـــة الجزائريـــة الديمقراطيـــة الشعبيـــة
People's Democratic Republic of Algeria

وزارة التعليـــم العالـــي والبحـــث العلمـــي
Ministry of Higher Education and Scientific Research

جامعـــة غـــردايـــــــة
University of Ghardaia

كـــليـة العـلـــوم والتكنولـوجيـــا
Faculty of Science and Technology

قســـم الرياضيـــات والاعـــلام الآلـــي
Department of Mathematics and Computer Science

ترخيص بمنـــاقشة مذكرة مـــاستر

Master Thesis Defense Permission

I, the undersigned, Dr. **Slimane Oulad-Naoui**

Hereby certify that I have examined the work entitled:

A multi source/task deep learning system for non-invasive detection and estimation of coronary artery stenosis

Presented to the partial fulfillment of the Master degree in Computer Science by:

- **Hocine Harrouzi**
- **Houdaifa Yahia-cherif**

I reviewed the document, and declare it is free from any serious defaults and respects the academic integrity rules. Furthermore, my jury member proposal is the following:

Slimane Bellaouar	MCA	Univ. Ghardaia	President
Youcef Madjoub	MAA	Univ. Ghardaia	Examiner
Slimane Oulad-Naoui	MCA	Univ. Ghardaia	Supervisor
Abderrahmane Cherif	PhD Student	Univ. Ghardaia	Co-Supervisor

Issued for all due intents and purposes.

Ghardaia, on June, 10, 2026 **Signature**

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة غرداية
كلية العلوم والتكنولوجيا
قسم الرياضيات والاعلام الالي

الرقم:...../ق.ر.ا/ك.ع.ت/ج.غ

شهادة الترخيص بالإيداع

أنا الأستاذ: سليمان بالاعور

المسؤول عن التصحيح مذكرة الماستر الموسومة بـــ

A multi-source/task deep learning system for non-invasive detection and estimation of coronary artery stenosis

من انجاز الطالب: حسين حروزي

والطالب: حذيفة يحي الشريف

شعبة: اعلام الي

تخصص: الانظمة الذكية لاستخراج المعارف

تاريخ المناقشة: 2026/06/21

أشهد ان الطلبة قد قاموا بالتعديلات والتصحيحات المطلوبة من طرف لجنة المناقشة، وأن المذكرة قد استوفت جميع شروطها وبإمكانهم إيداعها على مستوى مكتبة الكلية.

غرداية في .....

امضاء المسؤول عن التصحيح

سليمان بالاعور

رئيس القسم

رئيس قسم الرياضيات والإعلام الآلي
الحاج موسى ياسين