Impact of Connectivity-Preserving Loss Functions on the Segmentation of Thin Tubular Structures: Application to Coronary Arteries From CT Angiography Data

Master Thesis

submitted by

Denis Krnjaca, B.Sc.

Institute of Biomedical Engineering
Prof. Dr.-Ing. M. Francesca Spadea
Co-Referee: Ciro Benito Raggio, M.Sc.
Karlsruhe Institute of Technology
2026

Supervisors: Dr. Hannes Nickisch and Dr. Harald Hesse, Philips Hamburg.

Statutory declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Karlsruhe, March 11, 2026

Abstract

Coronary artery disease remains a leading cause of mortality worldwide. Coronary Computed Tomography Angiography (CCTA) provides a non-invasive basis for diagnosis; however, an accurate and connectivity-preserving segmentation of the coronary artery tree is essential for robust automatic and quantitative analyses. Convolutional Neural Networks (CNNs)-based architectures, in particular U-Net and no-new-U-Net (nnU-Net), have shown outstanding performance across a wide range of medical image segmentation benchmarks, yet they may frequently produce fragmented vessel trees when segmenting thin, tubular structures such as coronary arteries. Recent studies indicate that connectivity-aware loss functions can mitigate these discontinuities by explicitly penalizing missing centerline segments, but their efficacy for coronary artery tree segmentation remains to be demonstrated.

This thesis quantifies the benefits and challenges of integrating connectivity-preserving loss functions into an nnU-Net-based pipeline for one-step coronary artery tree segmentation from CCTA images. Performance is assessed using complementary metrics covering vessel mask accuracy, vessel accuracy, centerline completeness, and runtime, capturing volumetric overlap, connectivity-related effects, and computational cost. To contextualize the quantitative results, we conduct qualitative case studies with targeted visualizations.

The comparison of connectivity-preserving losses reveals that the Skeleton Recall (SR) loss provides the most consistent improvements in connectivity metrics while incurring substantially lower training time than the other connectivity-preserving loss formulations. In the final statistical analysis, augmenting the generic loss with a SR term improves coronary artery tree connectivity in a statistically significant and practically relevant manner, without substantially degrading volumetric overlap and with negligible computational overhead.

These findings identify SR as an effective, computationally efficient, and straightforward-to-implement loss function, making it a practically viable choice for accurate and connectivity-preserving coronary artery tree segmentation.

Contents

Abbreviations

CAD	Coronary Artery Disease 1, 8
CAD-RADS	Coronary Artery Disease Reporting and Data System 1
cbDice	centerline boundary Dice . . . . . 3 f., 29 ff., 33 f., 59
CCTA	Coronary Computed Tomography Angiography i, 1 f., 5, 7, 13 f., 16, 21, 23–27, 35 f., 38, 41, 45, 62 f., 74 f.
CE	Cross Entropy . . . . . 2 ff., 28, 32, 38, 59 f.
clCE	centerline Cross Entropy . . . . . 3 f., 29 f., 32 f., 59
clDice	centerline Dice . . . . . 2 ff., 29–33, 41, 43, 59
clTPR	centerline True Positive Rate . . . 43, 53 f., 56, 58, 64 f., 67–70, 80
CNN	Convolutional Neural Network . . . . . i, 1, 16, 35
CT	Computed Tomography . . . . . 5, 11 f., 24
DIU	Discrepancy between Intersection and Union 44, 53
DoG	Difference of Gaussian . . . . . 62 f., 80
ECG	Electrocardiogram 14, 78
FFR	Fractional Flow Reserve 1, 8
FN	False Negative . . . . . 40, 42, 45 f., 73, 75 f., 82
FP	False Positive . . . 27, 40, 42, 45, 52 f., 62 f., 66 ff., 73 f., 79 f., 82
GC	Gap Count . . . . . 43 f., 58, 70
HD	Hausdorff Distance 41, 52
HD95	95 % Hausdorff Distance . . . . . 2 f.
HU	Hounsfield Unit . . . . . 12 ff., 16, 25, 35 ff., 76, 81
LAD	Left Anterior Descending . . . . . 6 f., 73, 78
LCA	Left Coronary Artery 6
LCx	Left Circumflex . . . . . 6, 57 f., 78
LM	Left Main 6
MDCT	Multi Detector Computed Tomography 12
nnU-Net	no-new-U-Neti, 1, 5, 17 ff., 21, 34 ff., 38 ff., 49, 51, 54 ff., 66, 70, 79 f.
OM	Obtuse Marginal 7
PDA	Posterior Descending Artery 7
PLB	Posterior Lateral Branch 7
PPV	Positive Predictive Value . . . . . 62 f., 65
Q-Q	Quantil-Quantil 47, 51
RCA	Right Coronary Artery . . . . . 6 f., 57, 78

RI	Ramus Intermedius 6
ROI	Region of Interest 52, 66, 82
SA/V	Surface-Area-To-Volume Ratio 2
SCCT	Society of Cardiovascular Computed Tomography 7, 38
sMPR	stretched Multiplanar Reformat . . . . . 40, 45, 57, 71 ff.
SR	Skeleton Recall i, 2 ff., 29 f., 34, 38, 50, 52, 56, 58 f., 62–65, 71, 74, 78–81
TP	True Positive . . . . . 29, 31, 34, 42 f., 46, 65, 67, 72 f., 75 f.
TPR	True Positive Rate 65

Chapter 1

Introduction

1.1 Motivation

Coronary Artery Disease (CAD) continues to be a major cause of global morbidity and mortality [1]. This disease is characterized by atherosclerotic plaques that narrow the coronary lumen—a process known as stenosis—which can impair myocardial perfusion and lead to ischemia or infarction [2]. CCTA has emerged as a primary non-invasive imaging modality for assessing coronary anatomy and stenosis severity due to its high spatial resolution [3–5]. However, CCTA primarily provides an anatomical characterization and does not directly indicate hemodynamic significance. To standardize reporting and guide downstream management, Coronary Artery Disease Reporting and Data System (CAD-RADS) is used for patients undergoing CCTA [6]. To assess whether a stenosis is hemodynamically significant, additional functional evaluation is advised, such as CT Fractional Flow Reserve (FFR) [7], myocardial CT perfusion [8], or functional biomarkers [9]. Coronary artery segmentation is a key base technology for these subsequent analyses. Beyond diagnostics, accurate segmentations also enable virtually reconstructed coronary models for in-silico stent simulation and planning [10–12] and can support education and training [13, 14].

General-purpose CNN architectures, such as the U-Net [15] and the self-configuring nnU-Net [16], demonstrate outstanding performance across a wide range of anatomical structures. However, they often struggle to preserve topology in thin, branching anatomy. For coronary trees, structural integrity—particularly connectivity—is critical. Broken branches can conceal stenotic segments and distort downstream functional computations. In contrast, limited over- or under-segmentation of the vessel radius can often be tolerated, as it generally does not substantially affect subsequent functional analyses. To address this limitation, state-of-the-art methods often rely on complex, multi-stage pipelines that incorporate cascaded subnetworks [17, 18] or post-processing steps like centerline reconnection and mask reconstruction [19]. In such stage algorithmic pipelines, errors potentially accumulate and amplify and fixing of issues is cumbersome as there are multiple failure modes.

A plausible reason for the observed performance gaps in generic frameworks lies in the training objective itself. Generic losses (e.g., Dice, Cross Entropy (CE)) optimize volumetric overlap, where all voxels contribute equally to the loss, and therefore do not explicitly penalize missed voxels along centerlines. For structures with a high surface-area-to-volume ratio Surface-Area-To-Volume Ratio (SA/V), predictions are therefore prone to discontinuities. Recently, connectivity-aware losses, such as centerline Dice (clDice) [20] and the SR [21] have been proposed to address SA/V effects and tree topology, but their efficacy for coronary artery trees remains to be established.
Therefore, this work aims to systematically evaluate the benefits and challenges of integrating tailored loss functions into a generic CNN framework such as nnU-Net for the segmentation of coronary artery trees from CCTA images in a single analysis step. To this end, we conduct a comprehensive benchmark to assess performance not only in terms of traditional mask overlap but, crucially, with respect to topology-aware metrics. Our evaluation covers four complementary quality criteria: (a) centerline completeness, (b) vessel mask accuracy, (c) vessel accuracy, and (d) computational runtime.

1.2 State of the Art

Deep learning has transformed medical image segmentation, with U-Net [15] and nnU-Net [16] providing state-of-the-art baselines across diverse tasks. When sufficient curated data are available, these models facilitate accurate multi-structure segmentation in a single-stage setting [22]. Nevertheless, preserving connectivity in thin, curvilinear anatomy remains challenging for general-purpose tools. To better address such structures, various task-specific architectures have been proposed. These include attention-based designs tailored to capture long-range dependencies in curvilinear patterns [23] or vessel-specialized networks that jointly learn vessel masks, centerlines, and bifurcations [24]. Coronary segmentation is further complicated by the scarcity of publicly annotated cohorts, vessel diameters near the resolution limits of clinical CT, and high inter-patient variability in tree topology—especially in the presence of disease—which limits the utility of strong shape priors.
Community benchmarks reflect both progress and limitations. MICCAI organized challenges on centerline tracking (CAT, 2008) [25], stenosis assessment (STEN, 2012) [26], and full coronary tree segmentation (ASOCA, 2022) [27]. The ASOCA challenge is based on 40 CCTA images (20 normal, 20 diseased) with voxel-wise annotations [28]. These 40 cases served as the training set for ASOCA, complemented by an additional hidden test set of 20 CCTA images, which where provided without labels for submissions. In the ASOCA challenge, fully automated methods were primarily evaluated with Dice score and 95 % Hausdorff Distance (HD95). Top entries frequently relied on U-net or nnU-Net as a backbone within multi-stage pipelines and substantial post-processing to boost these metrics. After the official challenge phase, research has continued, and the public leaderboard remains active [29]. Rankings are listed under user handles and do not provide links to methods or public code, which limits attribution and comparability. As of Nov. 2025, the best result on the

public leaderboard is held by user “hongqq” with a Dice score of 0.8496 and a HD95 of 1.8879 mm. In the literature, Qiu et al. report leading results on the ASOCA training split using five-fold cross-validation (Dice 88.53%, HD95 1.07 mm) with a three-stage framework [19]. The pipeline first produces a voxel-wise coronary mask with nnU-Net. It then applies a regularized walk to reconnect broken centerlines by integrating distance, centerline-classifier probabilities, and directional cosine similarity. Finally, the pipeline employs an implicit neural representation to refine the geometry and recover missing vessel segments. However, a proper quantification of topological fidelity remains limited, as no explicit connectivity-aware metrics have been reported.

A complementary line of research modifies the training objective to directly favor topology. Early formulations inject topological priors using differentiable persistent homology [30] or enforce Betti-number constraints through a dedicated loss [31]. More recently, the use of a differentiable skeleton has gained traction with clDice, which introduced both a skeleton-based metric for topology-preserving similarity and a corresponding training loss that relies on a soft, differentiable approximation of the skeleton obtained via iterative min and max pooling [20]. In brief, clDice operates on skeletonized representations rather than pure volumetric overlap as in standard Dice, making the score sensitive to broken or missing vessel segments. Building on this idea, the centerline Cross Entropy (clCE) replaces the clDice ratio with CE terms, that are anchored to target and predicted skeletons. This change aims to improve robustness to noise and variability in annotations [32]. On the ASOCA dataset, the authors report 5-fold cross-validation with a Dice score of 84.80% and a clDice score of 84.95%. To jointly reflect topology and geometry, centerline boundary Dice (cbDice) augments clDice with boundary-aware terms and incorporates radius information from the distance transform along the skeleton. This approach addresses diameter imbalance while preserving connectivity [33]. Notably, cbDice leverages the topology-preserving differentiable skeletonization of Menten et al. [34], thereby promoting higher topological accuracy. A practical limitation of soft-skeleton based losses is the training-time overhead of computing differentiable skeletons. The SR loss, proposed by Kirchhoff et al. addresses this by precomputing a tubed skeleton from the ground truth and optimizing a soft recall of the prediction on that skeleton, thereby avoiding GPU-side differentiable skeletonization during training [21]. Across multiple public datasets of thin tubular structures, SR reports state-of-the-art performance while requiring only minimal additional training time and memory, making it an effective and computationally lightweight topology-aware training loss.

1.3 Research Question

Research question

What are the benefits and challenges of integrating connectivity-preserving losses into a standard architecture such as nnU-Net for coronary artery tree segmentation compared to using generic losses?

This thesis addresses the research question by systematically evaluating connectivity-preserving losses (SR, clDice, clCE, cbDice) against a generic baseline (Dice + CE) within the nnU-Net framework. We train and evaluate all models using a 5-fold cross-validation scheme on a combined dataset comprising the public ASOCA dataset and two in-house cohorts (98 volumes in total). Performance is assessed through a consistent protocol based on four complementary criteria: (a) centerline completeness, (b) vessel mask accuracy, (c) vessel accuracy, and (d) computational runtime. To isolate the contribution of the loss function and contextualize the quantitative results, we conduct targeted ablation experiments and qualitative visualizations. All experiments are performed on the Philips Innovative Technologies Hamburg compute cluster using an in-house PyTorch-based framework.

Chapter 2

Theoretical Fundamentals

In this chapter, the theoretical foundations required for this work are summarized. Following the structure of a typical medical image analysis pipeline, we begin by presenting the relevant medical background, then detail the acquisition of the imaging data, and finally describe their processing and segmentation.
Section 2.1 introduces the relevant medical background, including the anatomy of the heart and coronary arteries, as well as the pathophysiology of coronary artery disease, which provides the clinical context for cardiac imaging. Section 2.2 then reviews fundamental radiological principles, covering X-ray generation, Computed Tomography (CT), and specific aspects of CCTA acquisition. Section 2.3 formalizes the voxel-based representation of CCTA images and segmentation labels, discusses key image processing operations, and introduces the segmentation architectures employed in this thesis, namely the U-Net and the self-configuring nnU-Net framework.

2.1 Medical Fundamentals

2.1.1 Cardiac Anatomy

The heart is located in the middle mediastinum and functions as a dual pump that drives both pulmonary and systemic circulation. It consists of four chambers: the right and left atria, which function as receiving chambers, and the right and left ventricles, which act as ejecting chambers. A schematic overview of the cardiac chambers, valves, and great vessels is shown in Figure 2.1. Blood flow is regulated by the four cardiac valves. The atrioventricular valves comprise the tricuspid valve between the right atrium and the right ventricle and the mitral valve between the left atrium and the left ventricle. The semilunar valves are the pulmonary valve between the right ventricle and the pulmonary artery and the aortic valve between the left ventricle and the aorta [35].
The heart interfaces with the systemic and pulmonary circulation through the great vessels. The superior and inferior vena cava return systemic venous blood to the right atrium. The

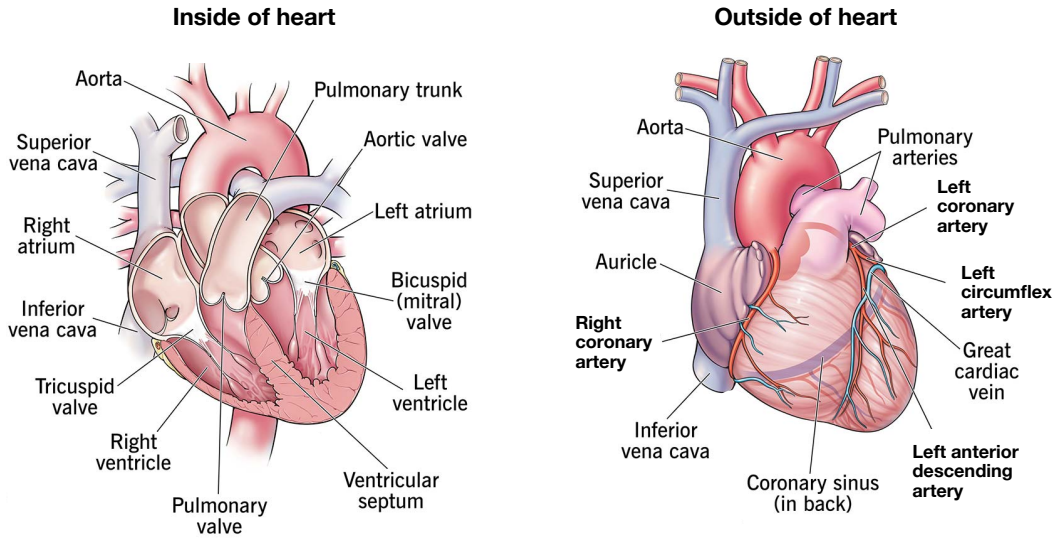

Figure 2.1: Anterior views of the human heart. Left panel: partially dissected heart. Right panel: external view. Both panels illustrate the main cardiac regions, heart valves, and major blood vessels; coronary arteries are labeled in bold. Adapted from [38].

pulmonary trunk bifurcates into the pulmonary arteries which supply the lungs. The pulmonary veins convey oxygenated blood to the left atrium, and the aorta carries oxygenated blood from the left ventricle into the systemic circulation [36].
The cardiac wall consists of three principal layers: the endocardium, which lines the cardiac chambers; the myocardium, which forms the contractile muscle layer; and the epicardium, which covers the outer surface of the heart and is closely related to the visceral layer of the pericardium [37]. The heart receives its blood supply via the coronary arteries, which run along the surface of the heart within the epicardial fat and provide branches penetrating the myocardium.

2.1.2 Coronary Artery Anatomy

The coronary arteries arise from the aortic root at the level of the right and left coronary sinuses. From the left coronary sinus, the Left Main (LM) coronary artery, also known as the Left Coronary Artery (LCA), originates, whereas the Right Coronary Artery (RCA) arises from the right coronary sinus [39].
The LM most commonly bifurcates into the Left Anterior Descending (LAD) artery and the Left Circumflex (LCx) artery; a trifurcation pattern with an additional Ramus Intermedius (RI) branch is a frequent anatomical variant [40]. The LAD artery courses within the anterior interventricular groove towards the cardiac apex and typically gives rise to septal perforator branches, which supply the interventricular septum, as well as and diagonal branches that supply the anterolateral wall. The LCx artery runs within the left atrioventricular groove and

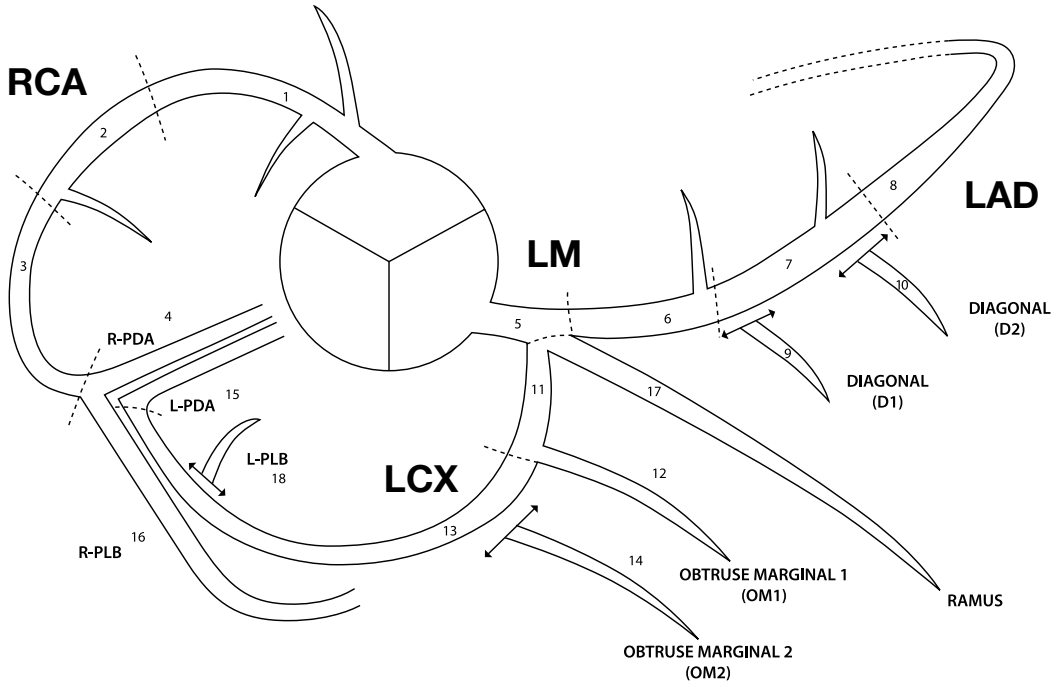

Figure 2.2: SCCT 18-segment coronary artery nomenclature adapted from [42]. Proximal (p), middle (m), distal (d). 1: pRCA, 2: mRCA, 3: dRCA, 4: right PDA, 5: LM, 6: pLAD, 7: mLAD, 8: dLAD, 9: D1, 10: D2, 11: pLCx, 12: M1, 13: mLCx, 14: M2, 15: left PDA, 16: right PLB, 17: ramus, 18: left PLB.

supplies the lateral and posterolateral walls of the left ventricle via Obtuse Marginal (OM) and Posterior Lateral Branch (PLB) [41].

The RCA follows the right atrioventricular groove and typically gives rise to the conus branch, sinoatrial nodal branch, and acute marginal branches. Distally, it gives rise to the Posterior Descending Artery (PDA) and PLB, depending on the coronary dominance pattern. Coronary dominance is defined by the vessel that gives rise to the PDA: approximately 80–85% of individuals exhibit right dominance, 15–20% left dominance, and about 5% co-dominance, although these proportions can vary across populations[40].

For standardized reporting in CCTA, the Society of Cardiovascular Computed Tomography (SCCT) recommends an 18-segment coronary artery model. This model assigns segment labels to proximal, mid, and distal portions of the major coronary branches and their side branches [42], as illustrated in Figure. 2.2. Common anatomical anomalies include separate conus origin, dual LAD, and anomalous coronary courses. Awareness of such variants is essential to avoid misinterpretation; however, a detailed discussion of coronary anomalies is beyond the scope of this work, and the reader is referred to [43] for a comprehensive overview.

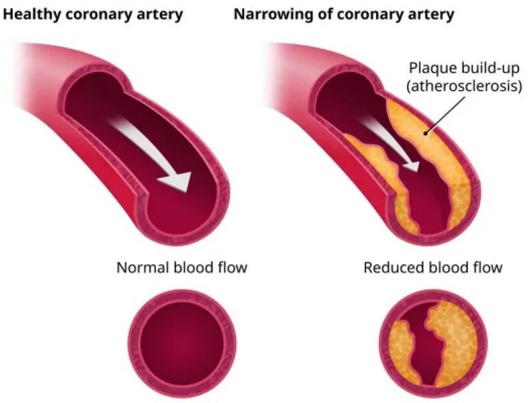

Figure 2.3: Schematic illustration of a normal coronary artery with unobstructed blood flow (left) and an artery with atherosclerotic plaque buildup causing luminal narrowing (right). Adapted from [47].

2.1.3 Coronary Artery Disease

CAD arises from the development of atherosclerotic plaque in the coronary arteries. Atherosclerosis is a chronic inflammatory process characterized by subendothelial retention of atherogenic lipoproteins, fibrous remodelling, and progressive calcification, ultimately leading to luminal narrowing, a process known as stenosis [44]. An illustrative depiction of normal and stenotic coronary arteries is shown in Fig. 2.3.

Plaque complications, particularly fibrous cap rupture or erosion with superimposed thrombosis, underlie acute coronary syndromes and myocardial infarction. Beyond these acute events, progressive luminal narrowing can become hemodynamically significant. A flow-limiting stenosis imposes an additional pressure drop across the lesion. Resistance vessels distal to the stenosis dilate to lower microvascular resistance and thereby preserve resting coronary blood flow via autoregulatory mechanisms. However, once maximal vasodilation is reached, hyperemic flow becomes limitted, coronary flow reserve declines, and a mismatch between myocardial oxygen supply and demand develops, resulting in myocardial ischemia [45].

The development and progression of CAD are driven by multiple, often interacting risk factors. Major causal contributors include hyperlipidemia, arterial hypertension, diabetes mellitus, cigarette smoking, and a sedentary lifestyle, among others [44]. The anatomical severity of a coronary stenosis does not always correspond to its hemodynamic significance. Lesion-specific ischemia is best assessed by physiological measurements such as FFR, and management decisions should integrate symptoms, ischemia burden, and overall cardiovascular risk [46].

For standardized anatomical and functional reporting in CCTA, CAD-RADS 2.0 provides a graded framework (categories 0–5 with additional modifiers) that harmonizes reporting and downstream management recommendations across centres [6].

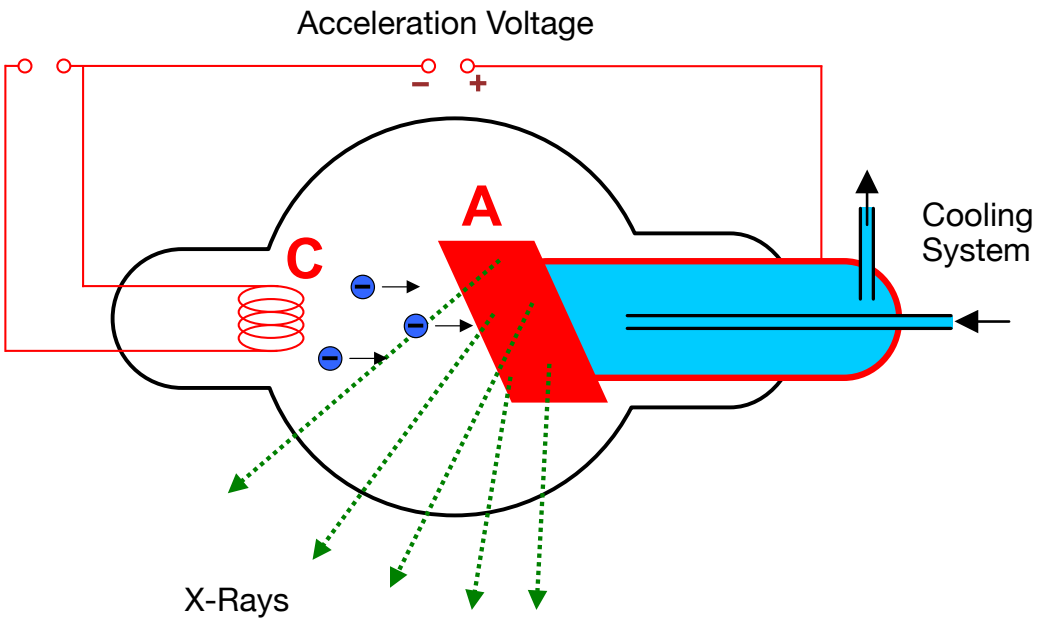

Figure 2.4: Schematic illustration of X-ray generation in an X-ray tube. Electrons are accelerated from the cathode (C) to the anode (A) and produce bremsstrahlung and characteristic X-ray photons upon deceleration in the anode material. Adapted from [50].

2.2 Radiology Fundamentals

2.2.1 X-rays

X-rays are a form of high-energy, ionizing electromagnetic radiation [48]. In medical imaging, they are typically generated in an X-ray tube, which consists of an evacuated glass or metal envelope containing a cathode and a solid metal anode [49]. Thermionic emission from a heated filament at the cathode releases electrons once the thermal energy exceeds the binding energy of the filament material. These electrons are then accelerated across a high potential difference between the negatively charged cathode and the positively charged anode. Upon impact with the anode, the fast electrons are decelerated and deflected in the electric field of the target atoms. This rapid deceleration of charged particles produces electromagnetic radiation, including X-rays. A schematic of X-ray generation in an X-ray tube is depicted in Fig. 2.4.

X-ray production in the anode material is dominated by two physical processes: bremsstrahlung and characteristic radiation [48]. In bremsstrahlung interactions, incident electrons are decelerated and deflected in the Coulomb field of the atomic nucleus, thereby losing part, and in extreme cases almost all of their kinetic energy. This energy loss is emitted as an X-ray photon. Because the fractional energy loss can vary continuously from nearly zero up to the full incident energy, bremsstrahlung gives rise to a continuous polyenergetic spectrum.

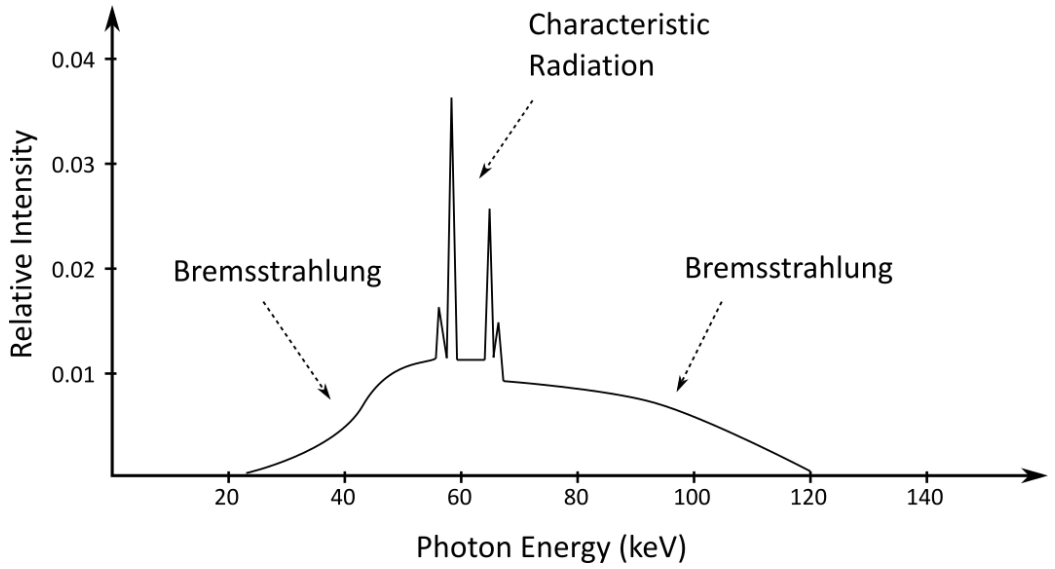

Figure 2.5: Example X-ray spectrum of a tungsten anode. Sharp peaks correspond to characteristic radiation, whereas the continuous background represents bremsstrahlung. Adapted from [49].

In characteristic X-ray production, an incident electron transfers sufficient energy to an inner-shell electron of a target atom to ionize it and create a vacancy. This "hole" is subsequently filled by an electron from an outer shell, and the energy difference between the two shells is emitted as an X-ray photon. The resulting spectrum consists of discrete lines at energies characteristic of the target material. An example spectrum of a tungsten X-ray tube illustrating the continuous bremsstrahlung background with superimposed characteristic peaks is shown in Fig. 2.5.

In a typical diagnostic X-ray tube, the vast majority of the electron energy is dissipated as heat through inelastic collisions with atomic electrons, whereas only a small fraction is converted into bremsstrahlung and characteristic X-rays [48].

When X-rays interact with matter in the diagnostic energy range, the two dominant processes in soft tissue and bone are the photoelectric effect and Compton scattering. In the photoelectric effect, an incident X-ray photon transfers all of its energy to a bound electron and is completely absorbed. The electron is ejected with kinetic energy equal to the photon energy minus the binding energy of the electron. This process is strongly dependent on the atomic number of the material and the photon energy, making it a major contributor to image contrast. Compton scattering describes an inelastic interaction between an X-ray photon and a weakly bound outer-shell electron. The photon transfers part of its energy to the electron, which is ejected as a so-called Compton electron, while the scattered photon emerges with reduced energy and is deflected from its original direction. Compton scatter contributes to image noise and reduces contrast because scattered photons may be detected outside their original projection path [51]. Schematic depictions of the photoelectric and Compton interactions are shown in Figure 2.6.

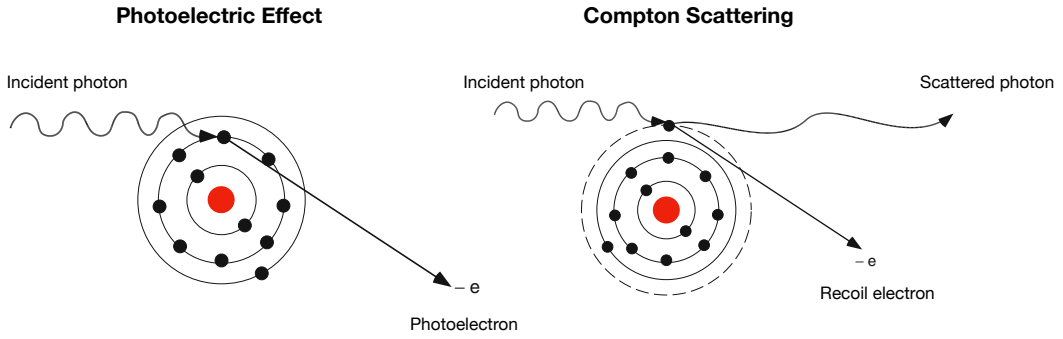

Figure 2.6: Schematic illustration of the two dominant processes for X-ray interaction with matter. Red circles indicate the atomic nucleus, and grey circles represent electrons in different shells. Left panel: schematic model of the photoelectric effect. Right panel: schematic model of Compton scattering. Adapted from [52].

The diagnostic utility of X-rays arises from their ability to penetrate tissue to a depth that depends on photon energy and material properties, thereby creating contrast between structures of different composition. By adjusting the tube voltage, one can modify the energy spectrum of the emitted X-rays and thus influence both image contrast and radiation dose. In a simplified, monoenergetic description, attenuation of X-rays in matter can be modeled by the Beer–Lambert law. When a monoenergetic X-ray beam of initial intensity I_0 traverses a homogeneous object with linear attenuation coefficient μ and intersection length x, the transmitted intensity I is given by

$$I = I_0\, e^{-\mu x}\,. \tag{2.1}$$

This model neglects scatter and the energy dependence of attenuation but forms the basis of the line-integral model used in CT reconstruction, in which attenuation is described as the integral of the linear attenuation coefficient accumulated along each X-ray path [53].

2.2.2 Computed Tomography

CT is an X-ray–based tomographic imaging technique in which cross-sectional images of the body are reconstructed from multiple X-ray projections acquired around the patient [54]. In conventional axial CT, a three-dimensional volume is obtained by acquiring, reconstructing, and stacking a series of two-dimensional slices along the axial (z) direction. Compared to conventional projection radiography, CT provides cross-sectional images that minimize superposition of overlying structures and therefore enable improved visualization of internal anatomy and pathology.

In a basic axial CT acquisition, often referred to as step-and-shot mode, the X-ray tube and detector rotate around the patient within the gantry while the patient table remains stationary for the duration of the rotation [55]. After each rotation, projection data for a given slice are reconstructed into a two-dimensional image. The table is then advanced by a predefined distance, and the process is repeated to acquire the next slice. For a given axial slice, all

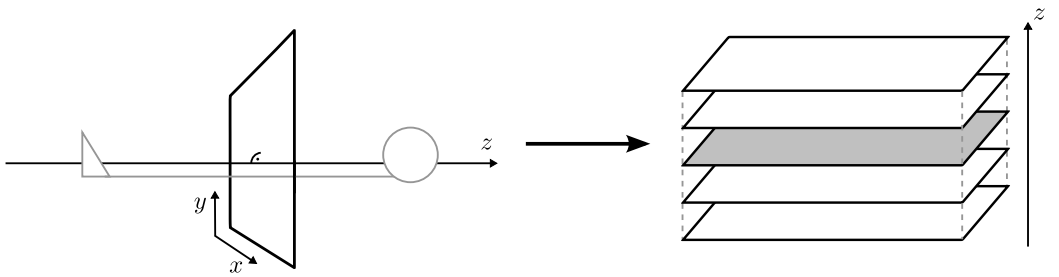

Figure 2.7: Conventional axial CT data representation. All projection rays for a given slice lie in a single x–y plane orthogonal to the beam direction z, so the object can be described by a bivariate function $f(x,y)$. A three-dimensional CT image is formed by stacking the reconstructed two-dimensional axial slices along the z-axis. Adapted from [55].

projection rays lie approximately within a single x–y plane orthogonal to the beam direction z, allowing the object to be modeled by a bivariate function $f(x,y)$ (Figure. 2.7).
The X-ray source trajectory and detector geometry are commonly described in terms of fan-beam geometry [55], in which all rays for a given projection angle emanate from a single focal spot and impinge on a curved or flat detector (Figure 2.8, left panel). Modern scanners typically employ Multi Detector Computed Tomography (MDCT) geometries [56]. Instead of a single row of detectors, multiple parallel detector rows are arranged along the z-axis, allowing simultaneous acquisition of several slices per rotation (Figure. 2.8, right panel). This increases longitudinal coverage per rotation and enables faster volume scanning. Even single-detector-row CT systems use multiple detector elements along the fan direction, but only a single detector row in the longitudinal direction. In MDCT, multiple slices are reconstructed from each projection set by exploiting the multiple detector rows, which is particularly advantageous for dynamic studies and angiographic applications.
For volumetric acquisitions, current CT systems typically employ helical CT, also referred to as spiral CT [57]. In helical CT, the X-ray tube rotates continuously around the patient while the table is translated at a constant speed along the z-axis, resulting in a helical source trajectory (Figure 2.9). Compared with step-and-shoot acquisitions, helical CT offers substantially faster coverage of the scan volume and reduces motion artifacts because the entire region of interest can be imaged in a single continuous run.
As outlined in Section 2.2.1, under the monoenergetic assumption and neglecting scatter, each X-ray measurement can be modelled by Beer–Lambert's law, which relates the measured intensity to the line integral of the linear attenuation coefficient along the corresponding ray [53]. In the context of CT, these measurements can therefore be interpreted as line integrals of $\mu(x,y,z)$ along a set of rays through the object. CT reconstruction algorithms (e.g. filtered backprojection or iterative methods) invert this transform to estimate the spatial distribution of μ in the scanned volume.
In clinical CT, reconstructed attenuation values are commonly mapped to the Hounsfield Unit (HU) scale, which is normalized such that water has a value of 0 HU and air approximately

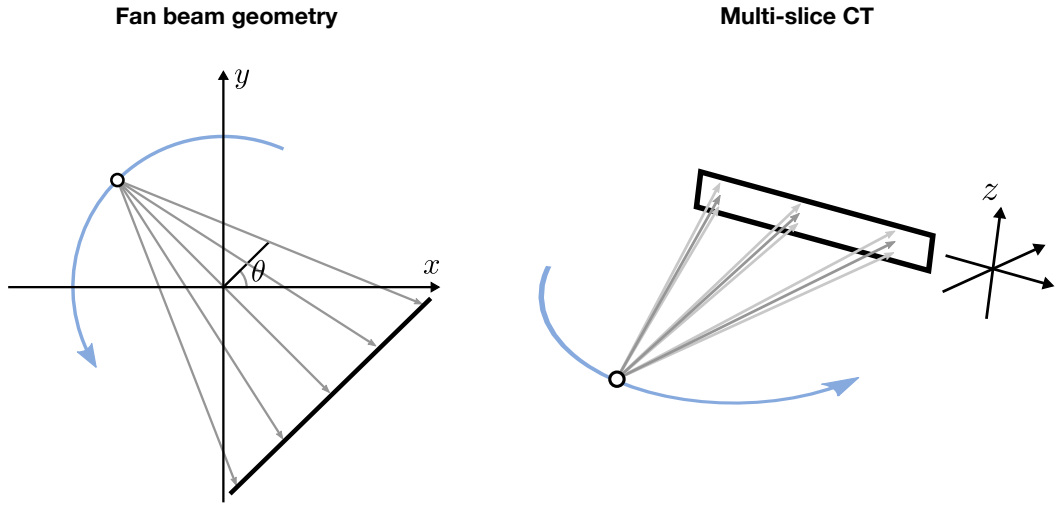

Figure 2.8: Basic acquisition geometries in CT imaging. Blue arrows indicate the trajectory of the X-ray source, and thick black lines depict the detector. Left panel: fan-beam geometry, where all rays for a given projection angle emanate from a single focal spot. Right panel: multi-slice CT geometry with multiple detector rows, enabling simultaneous acquisition of several image slices from one X-ray source position. Adapted from [55].

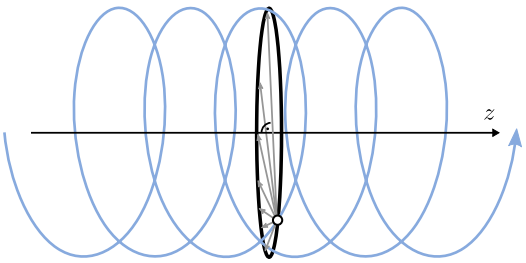

Figure 2.9: Helical CT acquisition geometry. The X-ray source rotates in the x–y plane while the patient table is translated along the z-axis, resulting in a helical source trajectory. Fan-beam projections for a given axial position can be obtained by interpolation between neighboring points on the helix.

$-1000\,\mathrm{HU}$. The HU of a given material is defined as

$$\mathrm{HU} = 1000 \cdot \frac{\mu_{\mathrm{material}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}}, \tag{2.2}$$

where μ_{material} and μ_{water} denote the linear attenuation coefficients of the material and water, respectively[58]. Typical HU values for relevant tissues and materials are summarized in Table 2.1.

2.2.3 Coronary Computed Tomography Angiography

Cardiac CT, and CCTA in particular, poses specific challenges due to the continuous motion of the heart. The key technical requirement is sufficient temporal resolution to effectively

Table 2.1: Typical HU values and ranges for different tissues and materials. Exact values depend on tissue composition, tube voltage, and temperature [59–61].

Material / Tissue	HU
Air	-1000
Fat	-100 to -80
Water	0
Muscle	$+10$ to $+40$
Blood	$+30$ to $+45$
Soft tissue	$+40$ to $+80$
Opacified blood	$+250$ to $+350$
Bone	$+400$ to $+2500$

"freeze" cardiac motion. If the acquisition is too slow or occurs at an unfavorable phase of the cardiac cycle, residual motion can leading to image blurring and impaired diagnostic quality [62].

Patient preparation is therefore critical. The heart rate is typically lowered to ≤ 60–$65\,\text{bpm}$ using oral or intravenous β-blockers [63]. In addition, an iodinated contrast bolus (e.g. Omnipaque) is administered via an intravenous line to opacify the coronary arteries, thereby increasing attenuation of the intravascular blood pool and improving image contrast [27]. CCTA acquisition is usually initiated using bolus tracking: scanning starts once the attenuation in a predefined region of interest, typically the ascending aorta, reaches a predetermined HU threshold [64].

Besides high temporal resolution, correct timing within the cardiac cycle is essential. Image acquisition is synchronized to the Electrocardiogram (ECG) using ECG gating. Two main approaches are employed: prospective ECG triggering and retrospective ECG gating. In prospective ECG triggering, data are acquired only during selected phases of the cardiac cycle, typically mid-diastole, when cardiac motion is minimal. The scanner monitors the ECG and initiates X-ray exposure after a specified delay following the R-wave. Between these acquisition windows, the X-ray tube is switched off, resulting in a step-and-shoot–like acquisition along the z-axis.

In retrospective ECG gating, the X-ray tube operates continuously with tube current modulation, and data are acquired over several cardiac cycles while the table moves through the gantry [62]. The ECG signal is recorded simultaneously, and images are reconstructed retrospectively at selected phases of the cardiac cycle, such as during systole or diastole. This approach enables functional assessment of ventricular performance and valve motion and allows flexible selection of phases with minimal coronary motion. However, it usually entails a higher radiation dose than purely prospective protocols [62].

A schematic overview of a typical CCTA workflow is shown in Figure 2.10.

Following these acquisition steps, the reconstructed images depict the coronary arteries in multiple anatomical planes. Figure 2.11 shows representative high-contrast CCTA images in axial, coronal, and sagittal views.

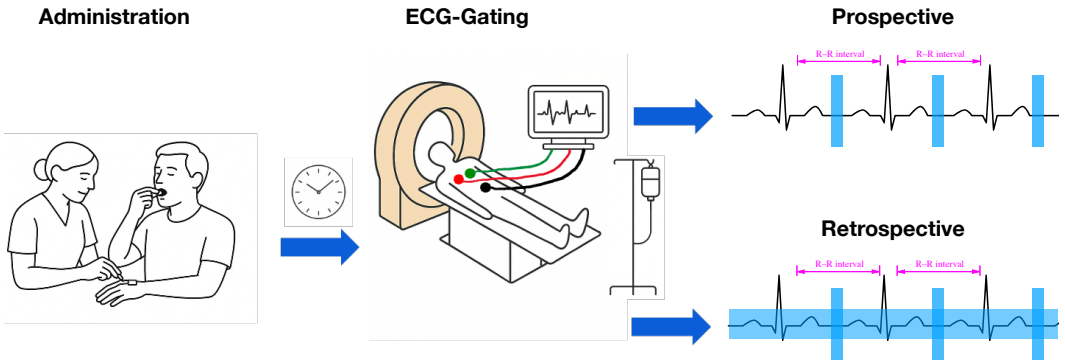

Figure 2.10: Workflow of a typical CCTA image acquisition. After administration of heart-rate-lowering medication, an intravenous contrast bolus is injected and bolus tracking is used to select the optimal scan start time. The patient is connected to ECG monitoring and the scan is synchronized to the R–R intervals. Data acquisition is performed under ECG gating either with prospective triggering (shaded blue intervals indicate periods when the X-ray tube is on) or with retrospective gating (tall shaded blue boxes indicate diastolic reconstruction windows; the continuous blue line illustrates continuous, ECG-modulated tube current).

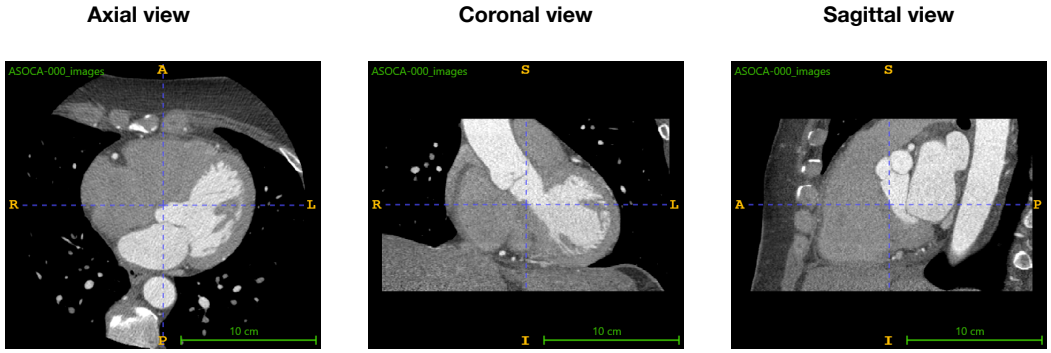

Figure 2.11: Example CCTA images in axial, coronal, and sagittal views illustrating the coronary arteries and cardiac anatomy.

2.3 Image Processing

2.3.1 Voxel-based Image Representation

In digital image processing, images are represented by numerical values associated with positions on a regular grid. In two dimensions, a single grid location is referred to as a pixel, and the associated numerical value is typically given as a gray value or colour channel intensity [65].

As introduced in Section 2.2.3, this work deals with three-dimensional images. In this setting, the numerical value is associated with a small volume element in 3D space, referred to as a voxel. A voxel-based image can therefore be regarded as a discrete mathematical function

that maps a grid position in three-dimensional space to a scalar value. Formally, the image is represented as

$$I[i, j, k] = p\,, \tag{2.3}$$

where i, j, k denote the voxel indices along the three image axes and p is a scalar intensity value. In CCTA imaging, p corresponds to the voxel's attenuation expressed in HU, as introduced in Section 2.2.2.

In addition to intensity images I, segmentation labels are represented as voxel-based images to serve as ground-truth annotations for training and evaluation of segmentation models. A label image

$$L[i, j, k] \in \{0, 1, \ldots, C-1\} \tag{2.4}$$

is defined on the same voxel grid as $I[i, j, k]$ and assigns a discrete class label to each voxel. In the simplest case of binary vessel segmentation used in this work, $C = 2$ and $L[i, j, k] \in \{0, 1\}$ indicate background and coronary artery voxels, respectively.

Following the conventions of SimpleITK [66], each image is characterised by an origin, specifying the physical location of the first voxel, a voxel spacing, defining the distance between neighbouring voxels along each axis, a size, giving the number of voxels in each dimension, and a direction cosine matrix, which describes the orientation of the image axes [67]. In CCTA images, the voxel grid is generally anisotropic. The in-plane resolution in x and y is often relatively homogeneous across scans, whereas the slice spacing in z shows substantially more variability between protocols. This results in elongated voxels along the through-plane direction and heterogeneous anisotropy across the dataset.

CNNs operate purely on the discrete voxel grid and do not natively account for differences in physical spacing, so all volumes are typically resampled to a common, near-isotropic target spacing before training [68]. For images, linear interpolation is commonly used for resampling intensity images, whereas nearest-neighbour interpolation is employed for label images to avoid generating mixed or fractional class labels. On a regular grid, each target voxel inside a grid cell is assigned a weighted average of the surrounding source voxels, with weights proportional to the relative distances along each image axis so that voxels closer to the target position contribute more strongly than distant voxels. Figure 2.12 illustrates these concepts on a 2D grid.

2.3.2 U-Net

U-Net, a fully convolutional neural network introduced in 2015, represents a landmark architecture for biomedical image segmentation [15]. Its structure is characterized by a symmetric encoder–decoder layout with skip connections, as illustrated in Fig. 2.13, combining a contracting path that captures contextual information with an expanding path that enables precise localization. This design allows simultaneous exploitation of high-level semantic information and low-level spatial detail, which is crucial for medical image segmentation tasks [69].

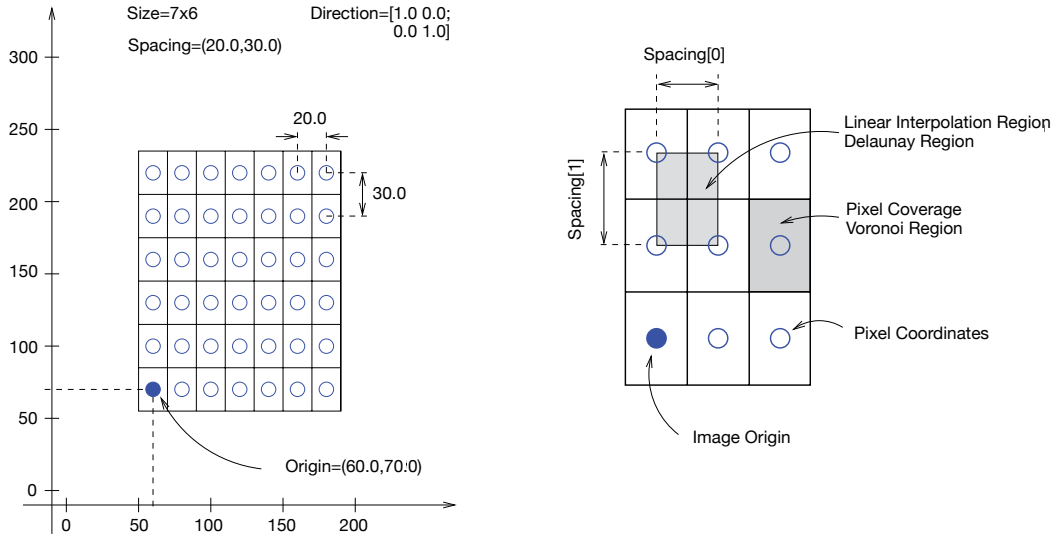

Figure 2.12: Illustration of pixel coordinates, spacing, and interpolation regions on a 2D image grid. Left panel: relationship between pixel indices and physical coordinates, including image origin and spacing. Right panel: pixel coverage (Voronoi region) corresponding to nearest-neighbour interpolation, where each point is mapped to the closest pixel centre, and linear interpolation regions (Delaunay regions), where the intensity is obtained as a continuous interpolation of several neighbouring pixels. Adapted from [66].

In the encoder, blocks of two successive convolutions, each followed by a nonlinear activation, are applied and then downsampled using max-pooling to halve the spatial resolution. With each downsampling step, the number of feature channels increases, allowing the network to encode increasingly abstract and semantically rich representations while reducing spatial dimensionality.
In the decoder, these operations are essentially reversed. Transposed convolutions are used to increase the spatial resolution and reduce the number of feature channels. At each resolution level, the upsampled features are concatenated with the correspondingly cropped feature maps from the contracting path via skip connections, followed by two convolutions with nonlinear activations. These skip connections help recover fine spatial details that might otherwise be lost during downsampling. Finally, a 1×1 convolution maps the multi-channel feature representation of the last decoder layer to the desired number of classes, producing a dense segmentation map. Since the output typically has the same spatial dimensions as the input image, a voxelwise loss between prediction and ground truth can be directly computed for training.

2.3.3 nnU-Net

The nnU-Net framework is a seminal advancement in medical image segmentation, proposing an automated and dataset-adaptive configuration of U-Net–based architectures [16]. Its central premise is that a well-configured, “vanilla” U-Net is difficult to outperform across

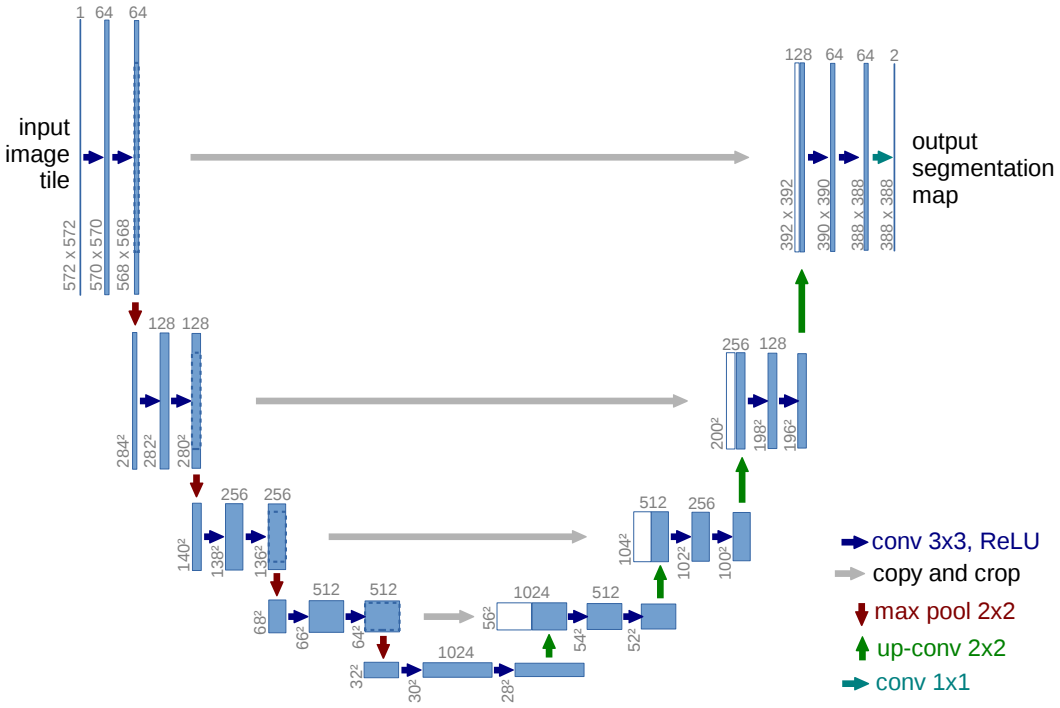

Figure 2.13: U-Net architecture (example for 32×32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box, and the x–y size is indicated at the lower left edge. White boxes represent copied feature maps passed via skip connections. Arrows denote the different operations (convolutions, pooling, upsampling). Taken from [15].

diverse segmentation tasks, and that much of the performance gap in practice arises from suboptimal, dataset-specific design choices rather than from architectural limitations [70]. Consequently, nnU-Net does not fundamentally change the network building blocks but automates the configuration and training pipeline that would otherwise require extensive manual tuning.

Conceptually, nnU-Net contrasts a typical U-Net workflow—where hyperparameters, preprocessing, and architecture variants are iteratively tuned for each new dataset—with an automated pipeline that infers most design choices from the data itself (Fig. 2.14).

The first step in the nnU-Net pipeline is to compute a dataset fingerprint that captures key dataset-specific properties, such as image size, voxel spacing, intensity distribution, imaging modality, number of classes, and number of training cases.

This dataset fingerprint is combined with a set of blueprint parameters, which are dataset-agnostic, and a set of inferred parameters derived from heuristic rules. Blueprint parameters include, for example, the loss function, optimizer, learning-rate schedule, and data augmentation strategy. Inferred parameters are determined from the dataset fingerprint via rule-based heuristics and include the intensity normalization scheme, resampling strategy for images and annotations, target spacing, network topology, and batch size.

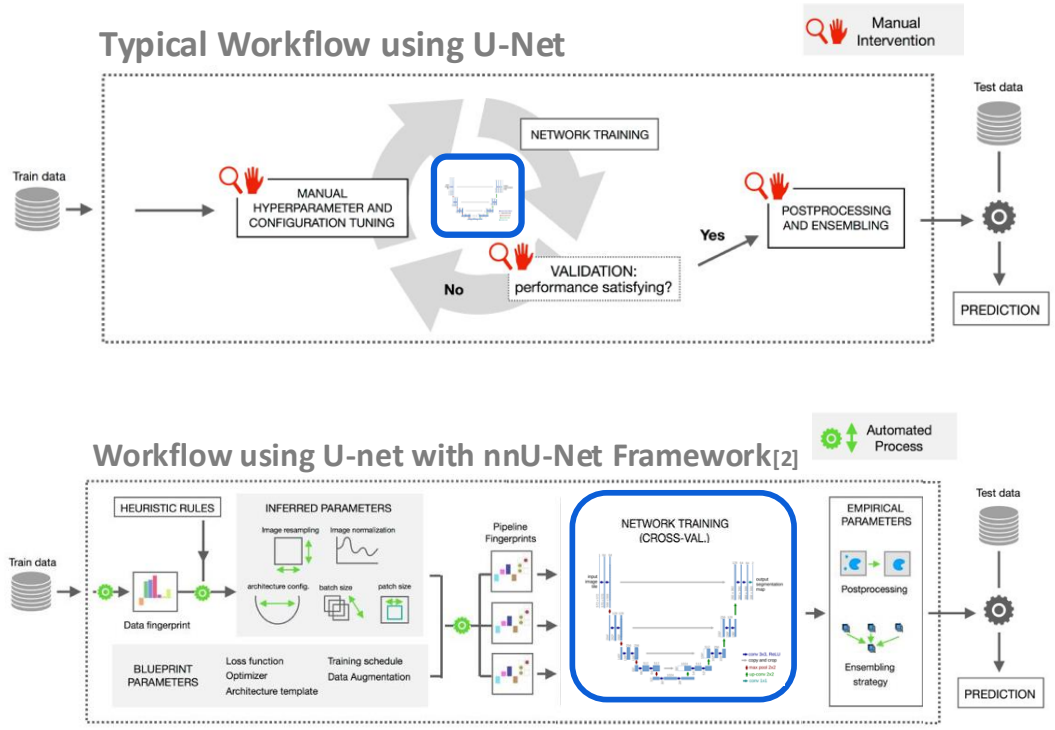

Figure 2.14: Comparison of a typical U-Net workflow and the nnU-Net framework. Upper panel: conventional workflow involving manual selection of hyperparameters and preprocessing for a specific dataset, training the model, evaluating its performance, and iteratively adjusting the configuration if the results are unsatisfactory. Lower panel: nnU-Net, built on a vanilla U-Net architecture, automates this loop by deriving most configuration choices from the dataset fingerprint, while only a small set of blueprint parameters needs to be specified by the user. Adapted from [68].

During training, nnU-Net employs deep supervision by attaching auxiliary loss functions to upsampled predictions from all decoder levels except the final one, thereby encouraging meaningful representations at multiple scales [71].

Chapter 3

Materials and Methods

This chapter describes the methodological and computational foundations of the proposed analysis pipeline, illustrated in Figure 3.1. Section 3.1 introduces the software environment used throughout this work. Section 3.2 details the composition annotation properties of the combined coronary CCTA cohorts and summarizes key image- and label statistics that inform subsequent preprocessing and network design. Section 3.3 formalizes the baseline generic loss and the evaluated connectivity-preserving losses, providing their mathematical definitions and implementation details. Section 3.4 describes the full training pipeline, including the initial exploratory setup and the configuration used within the nnU-Net framework. Section 3.5 defines the scoring metrics used to assess segmentation quality, describes the main visualization strategy employed for qualitative inspection, and outlines the methodology for the subsequent error analysis. Finally, Section 3.6 outlines the statistical analysis strategy used to assess distributional assumptions and select appropriate tests.

3.1 Tools and Framework

Python [72] served as the primary programming language for this work. All computations were executed on the Philips Innovative Technologies Hamburg cluster using OmniLearn[73]. For qualitative inspection of the CCTA volumes, coronary annotations, and segmentation results, we used ITK-SNAP [74] and the in-house Mirador viewer.

3.1.1 PyTorch

PyTorch [75] is an open-source deep learning library that provides tensor operations, automatic differentiation and efficient GPU acceleration. It is widely used in medical image analysis due to its flexibility, dynamic computation graph and extensive ecosystem for training convolutional neural networks.

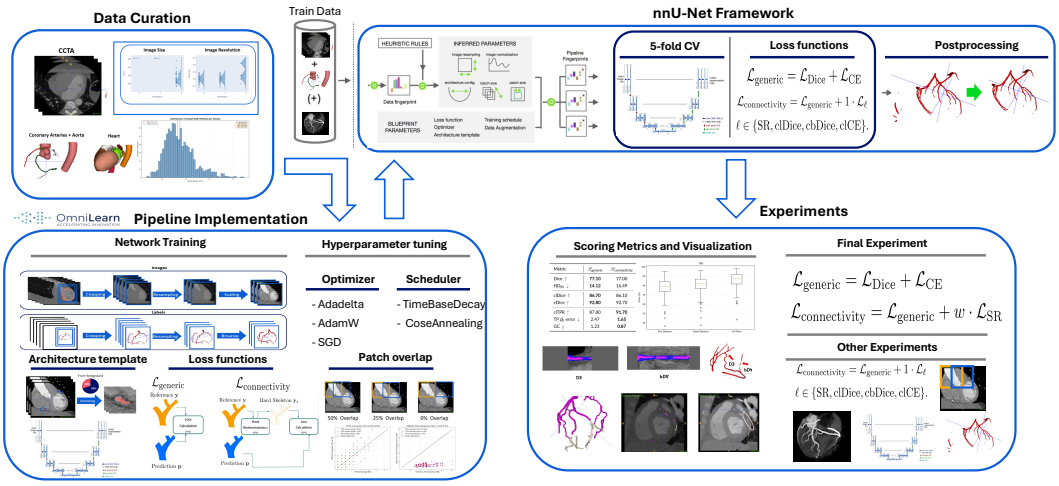

Figure 3.1: Overview of the analysis pipeline used in this work. The workflow begins with data curation, during which the available coronary CCTA datasets are inspected, validated, and refined where necessary. The curated data are then used in two successive training stages. First, an exploratory pipeline is employed to obtain an initial comparison between the generic loss and a connectivity-preserving loss and to tune the blueprint parameters. In the second stage, these parameters are fixed and an nnU-Net-based setup is trained to systematically compare the generic loss with several connectivity-preserving loss configurations. Segmentation performance is subsequently assessed using appropriate quantitative scoring metrics and qualitative inspection. In addition, a dedicated error analysis is conducted to identify typical failure modes, and statistical tests are employed to evaluate the significance of performance differences between the loss configurations. Figure for nnU-Net framework adapted from [68].

3.1.2 OmniLearn

OmniLearn is a Philips-internal deep learning framework for medical image analysis, built on top of PyTorch. Conceptually similar to MONAI [76], it provides modular components for data preprocessing, model training, inference, postprocessing and evaluation, and thereby covers the full AI cycle from raw data to deployable models. In addition to segmentation, OmniLearn supports tasks such as registration, landmark detection and classification, and includes infrastructure for experiment management and transfer of trained models into business and product environments. The overall framework structure is illustrated in Figure 3.2.

3.1.3 ITK-SNAP

ITK-SNAP is an interactive tool for displaying and annotating 3D medical images [74]. It supports efficient slice-based navigation, volume rendering and surface generation, making it suitable for inspecting anatomical structures and verifying segmentation quality.

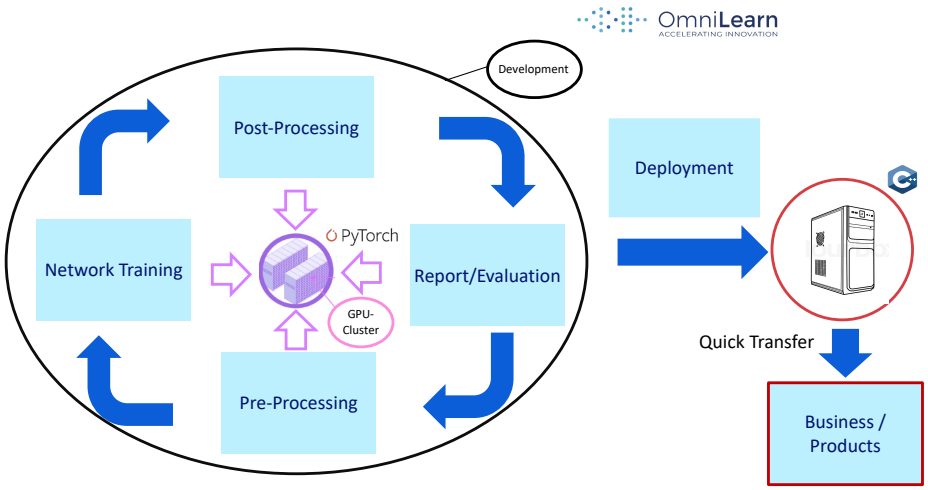

Figure 3.2: Schematic overview of the OmniLearn framework. It comprises modules for preprocessing, network training, postprocessing, evaluation and deployment, and is integrated with the internal GPU-cluster infrastructure for large-scale experiments.

3.1.4 Mirador

Mirador is an in-house visualization tool developed at Philips. It enables synchronized viewing of image volumes, ground-truth labels and segmentation predictions, which facilitates qualitative assessment of coronary artery connectivity and segmentation consistency.

3.2 Dataset

This section describes the CCTA datasets used in this work and their preparation for the segmentation pipeline. We first outline the composition of the combined cohorts and the available voxelwise annotations and then summarise key image and label properties that guide the choice of preprocessing settings and training parameters, before describing the curation steps applied to correct and refine the annotations.

3.2.1 Overview

The dataset used in this thesis consists of 98 CCTAs drawn from two sources: the public training dataset of the MICCAI 2020 ASOCA challenge and an in-house dataset comprising acquisitions from two clinical sites. The CCTA acquisition protocol underlying both cohorts is described in Section 2.2.3. Both datasets contain voxelwise multi-class annotations of

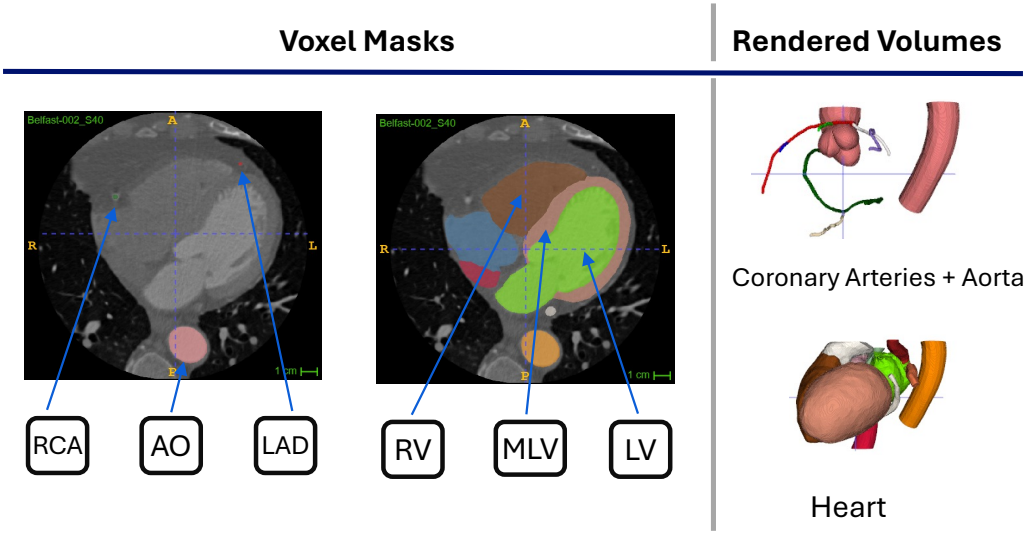

Figure 3.3: Visual examples of the voxelwise annotations available in the dataset. Left: axial CCTA slices with overlaid multi-class labels, including the RCA, LAD and aorta (AO) in the first image, and the right ventricle (RV), left ventricle (LV) and left ventricular myocardium (MLV) in the second image. Right: 3D volume renderings of the annotated coronary arteries, aorta and heart.

the coronary arteries and the aorta, as well as heart masks that delineate the cardiac region. Representative examples of these annotations are shown in Figure 3.3.
The ASOCA cohort comprises 40 CCTA scans, including 20 scans from healthy subjects and 20 scans from patients with confirmed coronary artery disease [28]. All scans were acquired on a GE LightSpeed 64-slice CT scanner under heart-rate control below 60 bpm. Coronary arteries were manually annotated by three expert readers, and a majority-vote fusion of their annotations was used to obtain the final reference segmentation.
The in-house cohort contributes 58 additional annotated CCTA scans acquired on Philips Brilliance 64 and Philips iCT 256 CT scanners under heart-rate control below 65 bpm. Twenty cases were reported as having no significant coronary artery stenosis, whereas the remaining 38 cases exhibited varying degrees of coronary artery disease.

3.2.2 Properties

In all cases, the in-plane matrix size is fixed at 512×512 voxels in the X- and Y-directions, consistent with conventional CT reconstruction, where 2D axial slices are acquired and stacked along the Z-direction (Fig. 2.7). The number of slices in the Z-direction varies substantially between patients, reflecting differences in the scanned anatomical range and acquisition protocol. The number of slices in the Z-direction varies substantially between patients, reflecting differences in the scanned anatomical range and acquisition protocol. The in-plane voxel spacings in X and Y show only minor variation across cases, whereas the spacing in Z exhibits markedly larger variability and is also coarser on average.

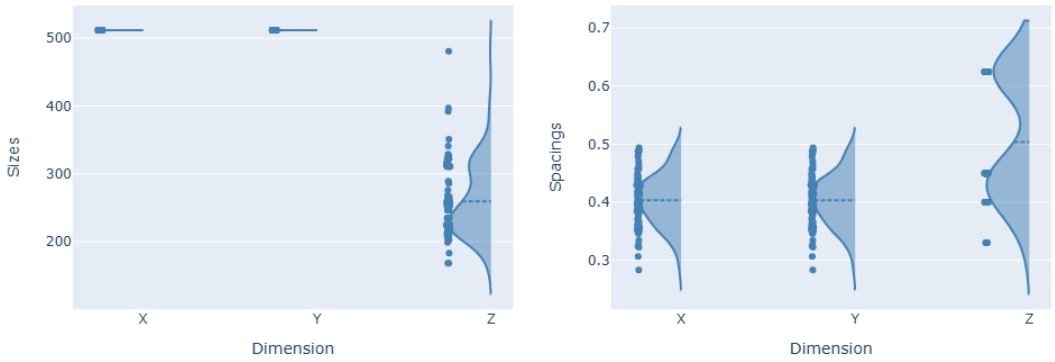

Figure 3.4: Distributions of image size and voxel spacing across the dataset. Left panel: distributions of image size in the X-, Y- and Z-directions. Right panel: distributions of voxel spacings in the X-, Y- and Z-direction.

	X	Y	Z
Min	512	512	168
Median	512	512	259
Max	512	512	481

Table 3.1: Image size statistics (in voxels) across the dataset. Here, X and Y denote the in-plane image width and height, and Z the number of slices.

	X	Y	Z
Min	0.283	0.283	0.330
Median	0.404	0.404	0.505
Max	0.494	0.494	0.625

Table 3.2: Image resolution statistics (voxel spacing in mm) across the dataset. Here, X and Y denote the in-plane spacing, and Z the slice thickness.

Figure 3.4 summarizes the corresponding distributions of image size and voxel spacing, whereas Tables 3.1 and 3.2 report the main summary statistics.

In contrast-enhanced CCTA, coronary arteries exhibit substantially higher intensities than the surrounding tissues due to intravascular iodinated contrast. Table 3.3 summarizes selected percentiles for both the full image and for voxels belonging to the annotated coronary arteries, illustrating the shift towards higher intensities in the contrast-enhanced vessels.

The overall image intensities span a wide range of HUs, reflecting the mixture of lung, soft tissue and bone, with a median of around -191HU. For the coronary foreground, the median intensity is about 306HU, and the central 99 % of foreground voxels range from about -96HU and 928HU. In both cases, the maximum intensities lie far above the 99.5th percentile, indicating a small number of extreme high-intensity voxels, for example due to metallic implants.

Percentile	Overall	Foreground
Min	-1024	-1001
0.5th	-1024	-96
Median	-191	306
99.5th	591	928
Max	3090	2817

Table 3.3: Summary of selected intensity percentiles (in HU) for the overall image and for foreground voxels belonging to the annotated coronary arteries.

	Foreground Ratio
min	0.0296
mean	0.0694
max	0.1371

Table 3.4: Foreground ratio statistics (percentage of voxels belonging to the annotated coronary arteries) across the dataset.

In terms of label distribution, the coronary arteries occupy only a very small fraction of the 3D volumes. Table 3.4 summarizes the foreground ratio across the dataset.

3.2.3 Curation

The raw datasets contained several imperfections that required manual curation before use in the segmentation pipeline. The main curation steps are summarized in Figure 3.5 and comprise heart mask correction, connected-component analysis of the coronary label maps, and the identification and correction of clearly too short vessel annotations.
Curation followed an iterative multi-step procedure. After each correction, cases were re-inspected and earlier steps were revisited if necessary, as changes in one step could reveal additional issues in another.
In one ASOCA case, the heart mask was undersegmented, missing parts of the right ventricle. The mask was therefore corrected and extended so that it tightly enclosed the cardiac chambers and proximal coronary arteries.
A connected-component analysis was performed on the coronary label maps to assess the topological consistency of the annotations. As described in Section 2.1.2, the coronary arteries form two main connected trees, originating from the left and right coronary ostia. In some cases, the label maps contained small isolated components or branches that were not connected to either of the major coronary trees. Such clearly implausible components were removed or reconnected to the appropriate main branch based on the underlying CCTA image. In contrast, two cases legitimately contained three connected components. Figure 3.6 illustrates such an example: in addition to the right and left coronary trees, a separate conus branch originates from a distinct third coronary ositum at the the right coronary sinus,

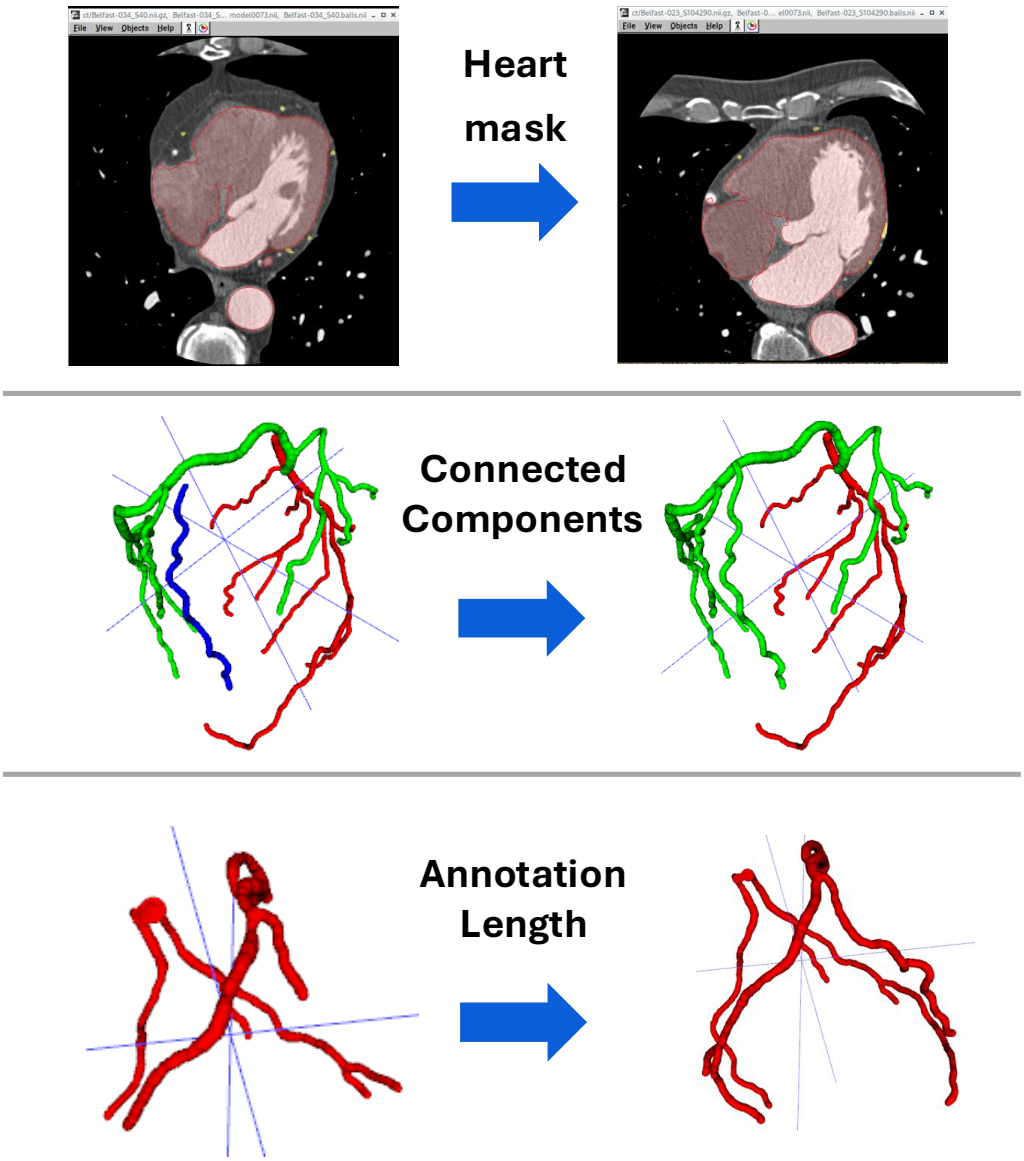

Figure 3.5: Overview of the main data curation steps. Top panel : correction of the heart mask to obtain an anatomically appropriate region of interest for cropping. Middle panel : connected-component analysis of the coronary label map to enforce anatomically plausible coronary trees. Bottom panel : correction of too short vessel annotations by extending distal segments.

representing a well-known anatomical variant rather than an annotation error. These cases were therefore retained unchanged.

The length and extent of the coronary annotations were inspected. In several cases, the vessels clearly continued distally in the CCTA images, while the corresponding labels stopped prematurely. Such truncated annotations artificially shorten the coronary trees and cause distally predicted vessel segments to be counted as apparent False Positives (FPs).

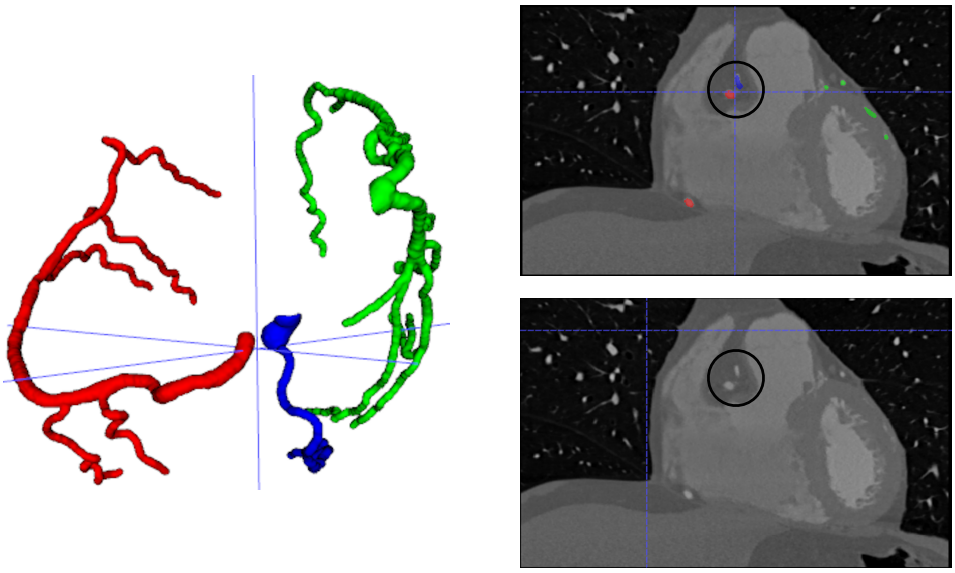

Figure 3.6: Example of a case with three connected components. Left panel: 3D rendering of the coronary label map showing the left coronary tree (green), the right coronary tree (red), and a separate conus branch (blue). Right panel: CCTA slices. The upper image shows the CCTA slice with the coronary labels overlaid, the lower image shows the raw CCTA slice. The black circle indicates the region of the separate conus ostium.

Where the distal vessel course was clearly identifiable, the annotations were manually extended to follow the contrast-enhanced lumen in the in-house cases.

3.3 Loss Functions

This section introduces the loss functions evaluated in this work. We first present the generic voxel-based losses that serve as our baseline, followed by four connectivity-preserving losses specifically designed for thin tubular structures.

3.3.1 Generic Losses

Segmentation networks for medical images are most commonly trained with generic loss functions that quantify voxel-wise agreement between the predicted segmentation and the ground-truth label. In this work, we consider the two most frequently used loss functions: the soft Dice and the soft CE.

3.3.1.1 Soft Dice

We define soft precision π and soft recall ρ as

$$\pi(\mathbf{y},\mathbf{p}) = \frac{\mathbf{y}^\top \mathbf{p}}{\mathbf{p}^\top \mathbf{1}}, \qquad \rho(\mathbf{y},\mathbf{p}) = \frac{\mathbf{y}^\top \mathbf{p}}{\mathbf{y}^\top \mathbf{1}}, \tag{3.1}$$

where $\mathbf{y} \in \{0,1\}^n$ denotes the binary reference mask and $\mathbf{p} \in [0,1]^n$ the corresponding prediction probabilities.
Here, $\mathbf{y}^\top \mathbf{p}$ represents the soft True Positives (TPs), i.e. the summed prediction probabilities at voxels that belong to the reference foreground. The denominator of π, $\mathbf{p}^\top \mathbf{1}$, is the sum of predicted foreground probabilities, so π measures which fraction of the predicted foreground lies inside the reference mask (soft precision). Conversely, the denominator of ρ, $\mathbf{y}^\top \mathbf{1}$, is the number of reference foreground voxels, so ρ measures which fraction of the reference foreground is captured by the prediction (soft recall).
Using these two terms, the soft Dice loss can be written as

$$\mathscr{L}_{\text{softDice}}(\mathbf{y},\mathbf{p}) = 1 - \frac{2}{\frac{1}{\pi(\mathbf{y},\mathbf{p})} + \frac{1}{\rho(\mathbf{y},\mathbf{p})}}, \tag{3.2}$$

which corresponds to one minus the soft Dice coefficient expressed as the harmonic mean of soft precision and soft recall.

3.3.1.2 Soft Cross-Entropy

For binary segmentation, the soft cross-entropy loss is defined via a per-voxel error vector

$$\mathbf{e}(\mathbf{y},\mathbf{p}) = -\,\mathbf{y} \odot \log(\mathbf{p}) - (\mathbf{1}-\mathbf{y}) \odot \log\big(\mathbf{1}-\mathbf{p}\big), \tag{3.3}$$

where $\mathbf{y} \in \{0,1\}^n$ denotes the binary reference mask, $\mathbf{p} \in [0,1]^n$ the corresponding prediction probabilities, $\mathbf{1}$ the all-ones vector, and $\odot$ denotes the Hadamard product. The soft cross-entropy loss is then obtained as the mean over all voxels,

$$\mathscr{L}_{\text{softCE}}(\mathbf{y},\mathbf{p}) = \frac{1}{n}\,\mathbf{1}^\top \mathbf{e}(\mathbf{y},\mathbf{p}). \tag{3.4}$$

Here, the first term in $\mathbf{e}(\mathbf{y},\mathbf{p})$ penalises foreground voxels with low predicted probabilities, while the second term penalises background voxels with low predicted background probabilities.

3.3.2 Connectivity-Preserving Losses

For thin branching structures such as coronary arteries, the most critical property of a good segmentation is the preserved connectedness of the vessel tree rather than absolute voxel-wise accuracy, see Figure 3.7. To explicitly encode connectivity during training, we consider four skeleton-based loss functions in this work: the SR [21], clDice [20], clCE [32] and cbDice [33].
For the SR (Section 3.3.2.4), we follow the original implementation and compute a binary (hard) skeleton using the 3D skeletonization routine from scikit-image [77], which uses

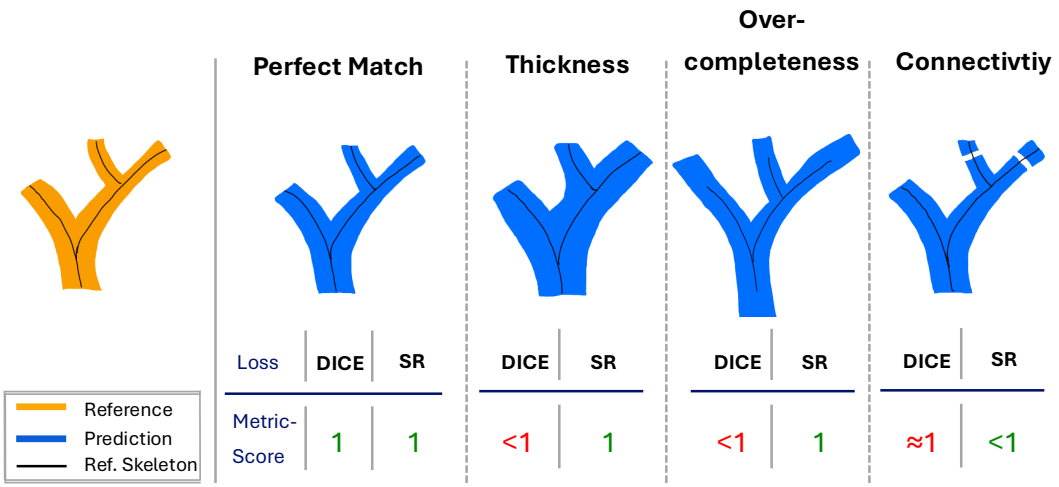

Figure 3.7: Illustration of how a generic loss (here Dice) and a connectivity-preserving loss (here SR) behave for different error types. The columns show four scenarios: perfect match, increased thickness, over-completeness and broken connectivity. In the thickness and over-completeness cases, the score corresponding to the connectivity-preserving loss remains 1, since the reference skeleton is fully recovered, whereas the score corresponding to the generic loss is reduced due to the additional volume. In the last column, a clear break in the vessel leads to almost no change in the score for the generic loss, because the overall volumetric overlap remains high, but the score for the connectivity-preserving loss drops markedly as parts of the reference skeleton are not recovered by the prediction.

an octree-based data structure to examine a $3 \times 3 \times 3$ neighbourhood of each voxel. The algorithm iteratively sweeps over the image and removes foreground voxels until the image stops changing. Each iteration consists of two steps: first, a list of candidate voxels for removal is assembled; second, voxels from this list are re-checked sequentially to better preserve the connectivity of the structure. Subsequently, we dilate the resulting skeleton with a diamond-shaped structuring element of radius 2 to obtain a tubular representation. A schematic overview of the SR computation is shown in Figure 3.8.

In contrast, clDice (Section 3.3.2.1, clCE (Section 3.3.2.2) and cbDice (Section 3.3.2.3) require a differentiable (soft) skeleton representation in order to be used as loss functions in gradient-based optimisation. For these losses, we employ the differentiable skeletonization method proposed in [34], which approximates the medial axis by a sequence of matrix additions and multiplications, convolutional operations, basic non-linear functions and sampling from a uniform probability distribution, all of which are fully compatible with backpropagation. A schematic overview of the computation for clDice and clCE is shown in Figure 3.9.

For cbDice, additional distance information is obtained using the GPU-based Euclidean distance transform routine from MONAI. Starting from the binary reference mask and the prediction, each foreground voxel is assigned its distance to the nearest boundary (background). After skeletonization, the distance values at skeleton voxels are interpreted as local vessel radius, yielding a radius-weighted skeleton. From these radius maps, normalized weight maps are derived that assign high weights to voxels close to the skeleton and lower

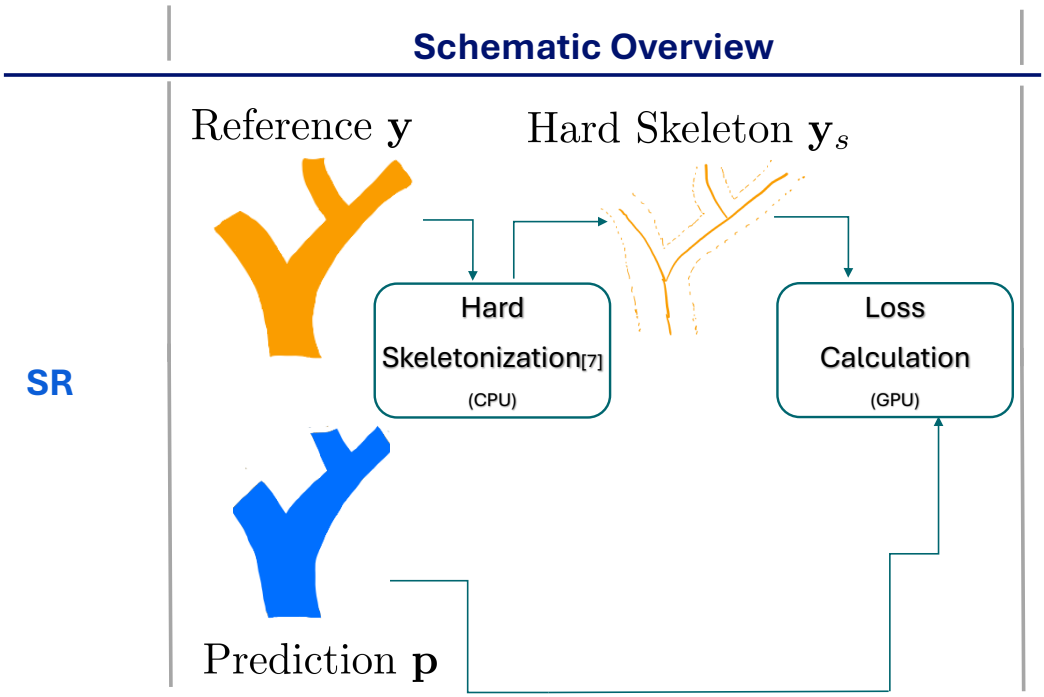

Figure 3.8: Schematic overview of the SR computation. The prediction mask $\mathbf{p}$ produced by the segmentation network is compared to the precomputed hard skeleton $\mathbf{y}_s$ of the reference mask $\mathbf{y}$ to obtain the SR loss term.

weights to voxels near the boundary. A schematic overview of the computation for cbDice is shown in Figure 3.10.

3.3.2.1 clDice

The clDice loss [20] was introduced as a loss function that leverages differentiable soft-skeleton representations, thereby increasing sensitivity to connectivity breaks compared to purely volumetric overlap losses for thin, tubular structures.
For its computation, we reuse the notion of soft precision and soft recall, but now defined on combinations of masks and soft skeletons. Let $\mathbf{y} \in \{0,1\}^n$ denote the binary reference mask, $\mathbf{p} \in [0,1]^n$ the prediction probabilities, and $\mathbf{y}_\sigma, \mathbf{p}_\sigma \in [0,1]^n$ the corresponding soft skeletons. We define

$$\pi(\mathbf{y}, \mathbf{p}_\sigma) = \frac{\mathbf{y}^\top \mathbf{p}_\sigma}{\mathbf{p}_\sigma^\top \mathbf{1}}, \qquad \rho(\mathbf{y}_\sigma, \mathbf{p}) = \frac{\mathbf{y}_\sigma^\top \mathbf{p}}{\mathbf{y}_\sigma^\top \mathbf{1}}. \tag{3.5}$$

Here, $\mathbf{y}^\top \mathbf{p}_\sigma$ represents the soft TPs of the predicted skeleton inside the reference foreground, i.e. the summed skeleton probabilities at voxels that belong to the reference mask. The denominator of π, $\mathbf{p}_\sigma^\top \mathbf{1}$, is the sum of predicted skeleton probabilities, so $\pi(\mathbf{y}, \mathbf{p}_\sigma)$ measures which fraction of the predicted skeleton lies inside the reference foreground (soft skeleton precision). Conversely, $\mathbf{y}_\sigma^\top \mathbf{p}$ quantifies how much of the reference skeleton is covered by

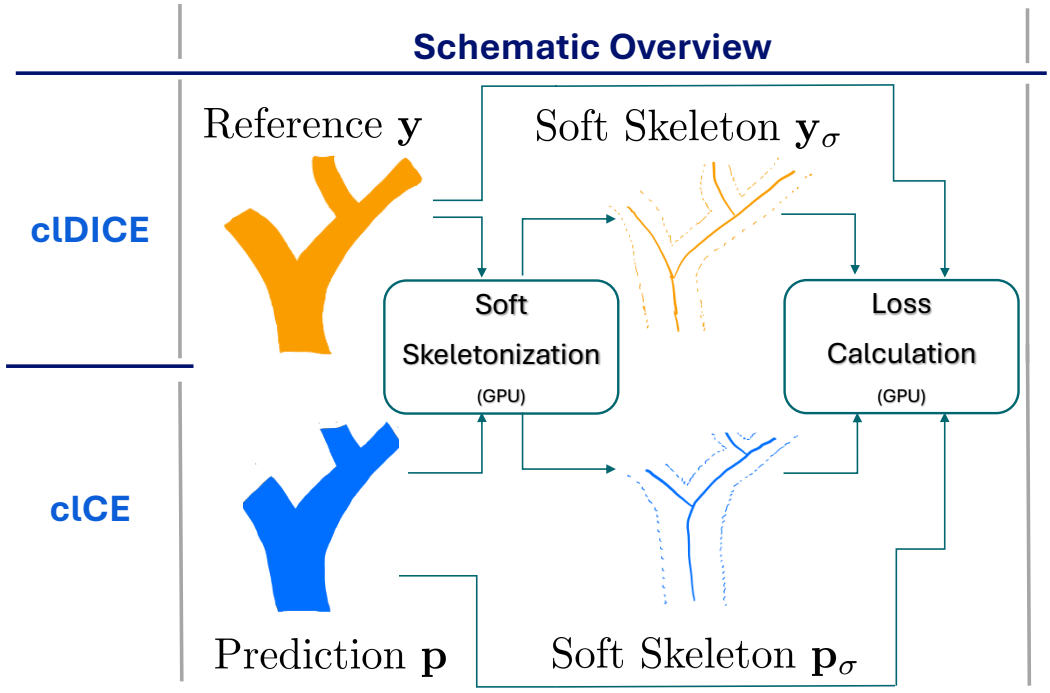

Figure 3.9: Schematic overview of the computation for clDice and clCE. The reference mask $\mathbf{y}$ and the prediction mask $\mathbf{p}$ are transformed into soft skeletons $\mathbf{y}_\sigma$ and $\mathbf{p}_\sigma$, respectively. These soft skeletons and masks are used to obtain the clDice loss term

the prediction, and $\mathbf{y}_\sigma^\top \mathbf{1}$ is the sum of reference skeleton probabilities, so $\rho(\mathbf{y}_\sigma, \mathbf{p})$ measures which fraction of the reference skeleton is recovered by the prediction (soft skeleton recall). Using these two terms, the clDice loss can be written as

$$\mathscr{L}_{\text{clDice}}(\mathbf{y}, \mathbf{p}, \mathbf{y}_\sigma, \mathbf{p}_\sigma) = 1 - \frac{2}{\frac{1}{\pi(\mathbf{y}, \mathbf{p}_\sigma)} + \frac{1}{\rho(\mathbf{y}_\sigma, \mathbf{p})}}, \tag{3.6}$$

which corresponds to one minus the clDice coefficient expressed as the harmonic mean of skeleton-based soft precision and soft recall.

3.3.2.2 Centerline Cross-Entropy

The clCE loss [32] was proposed to capitalize on the robustness of soft CE and the connectivity focus of the clDice loss, targeting improved overlap while maintaining faithful vessel network structure.
For its computation in binary segmentation, we first define the per-voxel CE error vector

$$\mathbf{e}(\mathbf{y}, \mathbf{p}) = -\mathbf{y} \odot \log(\mathbf{p}) - (\mathbf{1} - \mathbf{y}) \odot \log\big(\mathbf{1} - \mathbf{p}\big), \tag{3.7}$$

where $\mathbf{y} \in \{0, 1\}^n$ denotes the binary reference mask, $\mathbf{p} \in [0, 1]^n$ the corresponding prediction probabilities, $\mathbf{1}$ the all-ones vector, and $\odot$ denotes the Hadamard product.

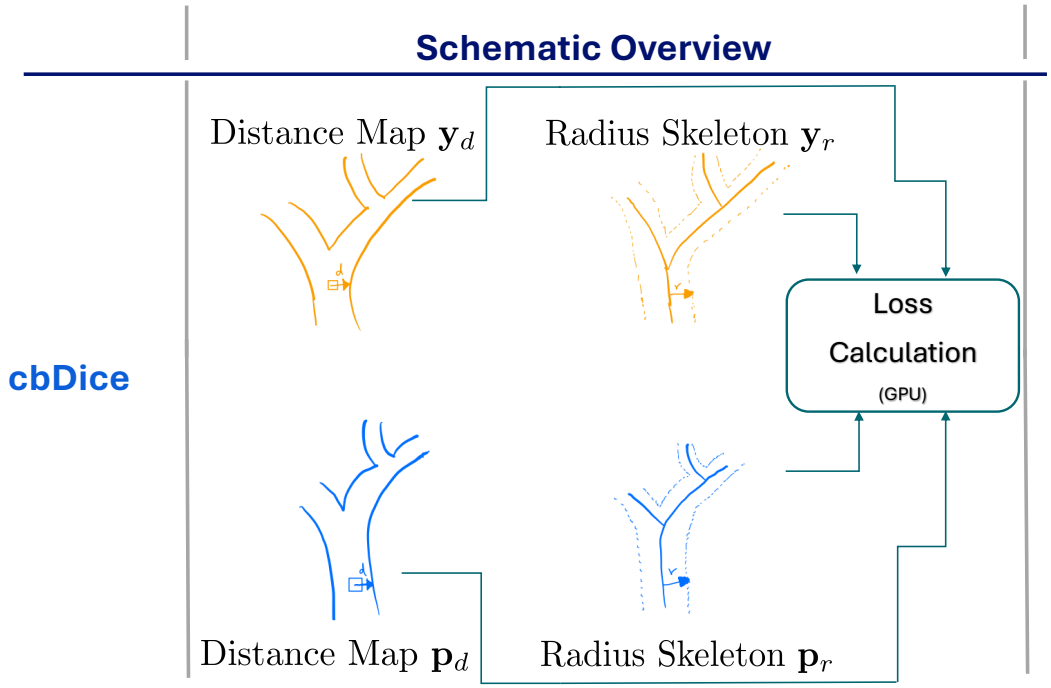

Figure 3.10: Schematic overview of the cbDice computation. For both reference and prediction, a Euclidean distance map ($\mathbf{y_d}$, $\mathbf{p_d}$) and a radius-weighted skeleton ($\mathbf{y_r}$, $\mathbf{p_r}$) are derived from the binary masks. cbDicecombines two topology-aware overlap terms: the predicted radius-weighted skeleton $\mathbf{p_r}$ is evaluated against the distance-weighted reference mask $\mathbf{y_d}$, and the reference radius-weighted skeleton $\mathbf{y_r}$ is evaluated against the distance-weighted prediction $\mathbf{p_d}$. The harmonic mean of these two terms yields the cbDice loss.

For clCE, this voxel-wise error is weighted by the soft skeletons of reference and prediction. Let $\mathbf{y}_\sigma, \mathbf{p}_\sigma \in [0,1]^n$ denote the differentiable soft skeletons of $\mathbf{y}$ and $\mathbf{p}$, respectively. The clCE loss is then defined as

$$\mathscr{L}_{\text{clCE}}(\mathbf{y}, \mathbf{p}, \mathbf{y}_\sigma, \mathbf{p}_\sigma) = \frac{1}{n}(\mathbf{y}_\sigma + \mathbf{p}_\sigma)^\top \mathbf{e}(\mathbf{y}, \mathbf{p}). \tag{3.8}$$

Here, $(\mathbf{y}_\sigma + \mathbf{p}_\sigma)$ acts as an importance-weighting map: voxels that lie close to the reference or predicted centerline receive higher weights, so their cross-entropy contribution dominates the loss, whereas voxels far from any skeleton have little influence.

3.3.2.3 Centerline Boundary Dice

The cbDice loss [33] extends the clDice formulation by incorporating boundary-aware and radius-related information, aiming to better capture geometric details.
For its computation, both the reference mask and the prediction are enriched with geometric information in the form of distance maps and radius-weighted skeletons. Let $\mathbf{y}_d, \mathbf{p}_d \in [0, d_{\max}]^n$ denote the Euclidean distance maps of the reference and prediction, assigning to each foreground voxel its distance to the nearest boundary. Likewise, let $\mathbf{y}_r, \mathbf{p}_r \in [0, r_{\max}]^n$

denote the corresponding radius-weighted skeletons, constructed by propagating the distance values onto the skeleton voxels. Using these representations, we define skeleton-based soft precision and soft recall as

$$\pi_r(\mathbf{y}_d, \mathbf{p}_r) = \frac{\mathbf{y}_d^\top \mathbf{p}_r}{\mathbf{p}_r^\top \mathbf{1}}, \qquad \rho_r(\mathbf{y}_r, \mathbf{p}_d) = \frac{\mathbf{y}_r^\top \mathbf{p}_d}{\mathbf{y}_r^\top \mathbf{1}}. \tag{3.9}$$

Here, $\mathbf{y}_d^\top \mathbf{p}_r$ measures how much of the predicted radius-weighted skeleton lies inside the distance-weighted reference mask (soft radius-skeleton precision), while $\mathbf{y}_r^\top \mathbf{p}_d$ measures how much of the reference radius skeleton is recovered by the prediction (soft radius-skeleton recall).
Using these two terms, the cbDice loss can be written as

$$\mathscr{L}_{\text{cbDice}}(\mathbf{y}_d, \mathbf{p}_d, \mathbf{y}_r, \mathbf{p}_r) = 1 - \frac{2}{\frac{1}{\pi_r(\mathbf{y}_d, \mathbf{p}_r)} + \frac{1}{\rho_r(\mathbf{y}_r, \mathbf{p}_d)}}. \tag{3.10}$$

which corresponds to one minus the cbDice coefficient expressed as the harmonic mean of skeleton-based soft precision and soft recall.

3.3.2.4 Skeleton Recall

The SR loss [21] was designed to preserve connectivity in thin tubular structures by maximizing recall on the reference skeleton. Since the skeleton is extracted once from the reference mask and treated as fixed during training, SR adds only a minor computational overhead compared with all other connectivity-preserving losses discussed in this section, which require differentiable skeletonization.
For its computation, let $\mathbf{y}_s \in \{0,1\}^n$ denote the binary skeleton extracted from the reference mask, and let $\mathbf{p} \in [0,1]^n$ be the predicted probability map. The SR loss is defined as

$$\mathscr{L}_{\text{SR}} = 1 - \frac{\mathbf{y}_s^\top \mathbf{p}}{\mathbf{y}_s^\top \mathbf{1}}. \tag{3.11}$$

The numerator $\mathbf{y}_s^\top \mathbf{p}$ corresponds to the soft TPs on the reference skeleton, i.e. the summed prediction probabilities at skeleton voxels. The denominator $\mathbf{y}_s^\top \mathbf{1}$ is the total number of skeleton voxels, so the fraction represents the proportion of the reference skeleton that is recovered by the prediction.

3.4 Training Pipeline

This section describes the training pipelines used in this work. The methodology is centered on a nnU-Net-based evaluation pipeline, which is used to obtain the primary results reported in this study. In addition, a lightweight exploratory pipeline was designed to enable rapid iteration and low-overhead preliminary experiments, thereby validating key design choices before establishing the nnU-Net-based evaluation setup.
The two pipelines share the same conceptual structure—preprocessing, network training, inference, and postprocessing—but differ in implementation details and computational cost.

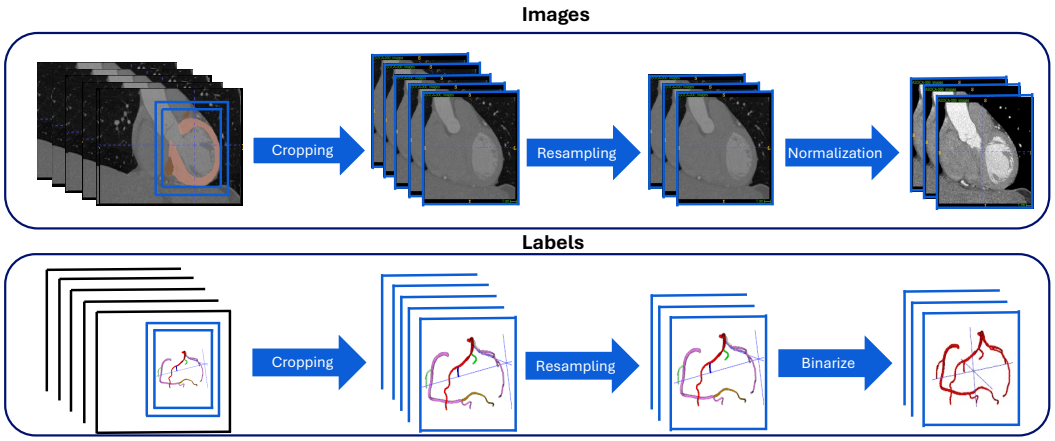

Figure 3.11: Preprocessing workflow of the exploratory pipeline. The figure shows two parallel processing streams for the CCTA images (top) and the coronary label masks (bottom). Both streams begin with a cropping step, where the volumes are restricted to a heart-centered Region of Interest (ROI). The cropped volumes are then resampled to a common target spacing to enable the network to learn spatial semantics consistently across cases. For the images, intensities are subsequently clipped to a predefined HU window and normalized, while the multi-class coronary annotations are converted into a binary mask representing foreground versus background.

3.4.1 Preprocessing

Preprocessing comprised of three main steps applied to the CCTA images and their corresponding label mask. The overall preprocessing workflow is summarized in Figure 3.11.

3.4.1.1 Cropping

nnU-Net performs cropping to the region of non-zero values. For our dataset, this operation has no effect, because the background in the CCTA images corresponds to air rather than water and therefore does not have a HU value of zero (Table 2.1).
To reduce computational cost in selected experiments and to enable faster turnaround during ablations, we therefore introduced an explicit, anatomically motivated cropping step based on the available heart mask.
For each case, the CCTA image together with its label mask was cropped by computing a bounding box around the union of the left ventricular myocardium and right ventricle labels derived from the heart mask. This bounding box was then isotropically expanded by 40 voxels in all directions to ensure that the full course of the coronary artery labels was contained within the cropped volume. To provide sufficient spatial context at the crop borders and reduce boundary artifacts during CNN training, an additional margin of 20 voxels was added prior to extracting the final crop. The overall procedure is illustrated in Figure 3.12.

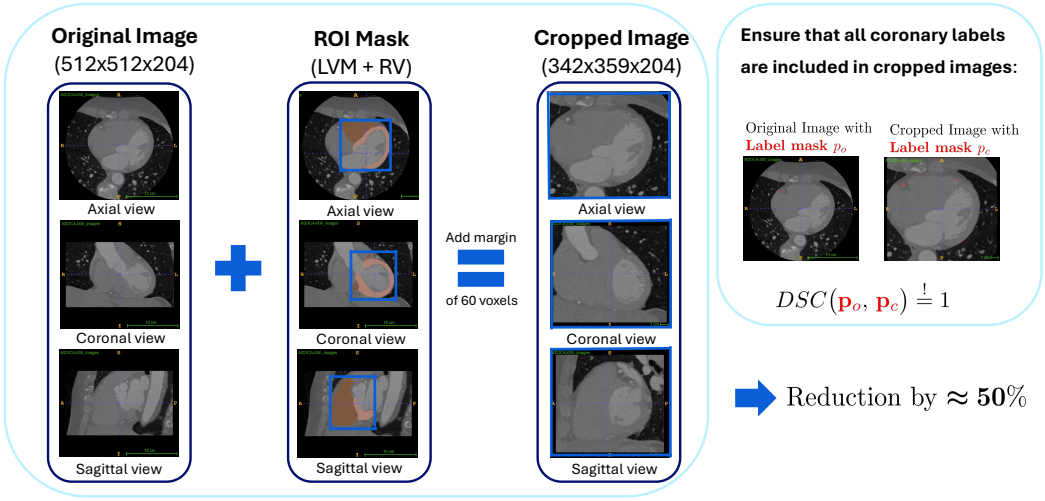

Figure 3.12: Visual illustration of the cropping workflow for a representative case. The left panel of the figure consists of three columns showing, from left to right, the original CCTA image volume (512×512×204), the Region of Interest (ROI) mask obtained, and the resulting cropped volume (342×359×204), each displayed in axial, coronal, and sagittal views. The right panel of the figure demonstrates that all coronary artery labels remain fully included in the cropped volume, which is quantitatively confirmed by a Dice score of 1 between the original and cropped label masks. Overall, this procedure reduces the image size by approximately 50%.

3.4.1.2 Resampling

nnU-Net determines a dataset-level target spacing by taking the per-axis median spacing across the training set and resamples all images accordingly. For image data, third-order spline interpolation is used, while label masks are resampled using nearest-neighbour interpolation.

In our dataset, in-plane spacing is already tightly clustered around 0.4 mm, whereas variability and coarser resolution primarily occur along the through-plane (Z) direction (Section 3.2.2). We therefore resample only along the Z-axis to 0.4 mm to standardize through-plane resolution while preserving the native in-plane sampling. For the CCTA images, we use linear interpolation, whereas the label masks are resampled using nearest-neighbour interpolation in the exploratory pipeline. Compared to the third-order spline interpolation used in nnU-Net, linear interpolation is computationally less expensive.

3.4.1.3 Normalization

nnU-Net collects foreground intensities across the training set and clips them to the [0.5th, 99.5th] percentiles of this distribution. In our dataset, this corresponds to a clipping range of $[-96, 928]$ HU. Subsequently, the clipped intensities are normalized using a z-score based on the mean and standard deviation of the same foreground distribution.

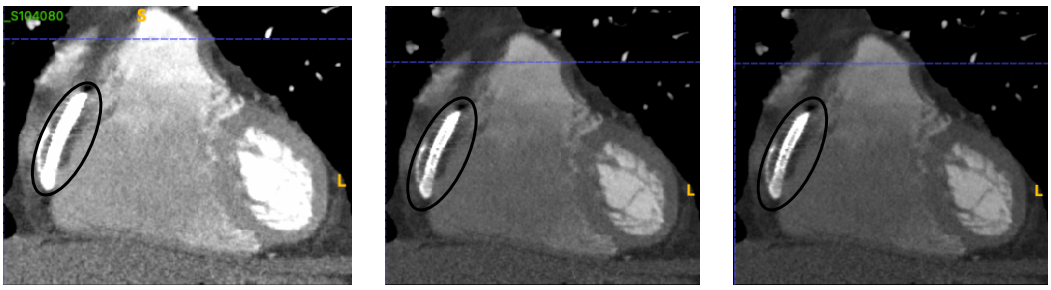

Figure 3.13: Effect of different upper clipping limits on the appearance of a contrast-enhanced coronary segment, highlighted with a black ellipse, in a representative CCTA case. From left to right, the same slice is shown with upper bounds of 500 HU, 1070 HU, and 1311 HU (lower bound fixed at -330 HU). With 500 HU, the contrast-enhanced lumen is heavily saturated and appears almost uniformly white. Extending the window to 1070 HU reduces saturation but still leaves parts of the lumen and calcifications fully clipped. Using the upper bound of 1311 HU preserves most of the intensity variation within the coronary lumen, although small regions of extremely high attenuation may still reach the upper limit and appear saturated.

For intensity normalization, we aimed in our exploratory pipeline to preserve the dynamic range of high-attenuation coronary voxels while suppressing clearly irrelevant extremes. A commonly used cardiac CT window is $[-330, 500]$ HU. However, as shown in Table 3.3, the 99.5th percentile of the *coronary foreground* intensities is substantially higher than 500 HU, so this range would clip a considerable fraction of coronary voxels.
As a clinical reference, we therefore considered a cardiac CT viewing preset in the Philips Advanced Visualization Workstation, which uses an upper window limit of 1070 HU. Even with this extended range, any voxel with an attenuation above 1070 HU is clipped to that value, causing extremely dense structures—such as heavily opacified coronary segments or calcifications—to appear uniformly white. As a result, relevant intensity differences in the high-attenuation range are lost.
We therefore adopted a data-driven approach: for each case, we computed the 99.5th intensity percentile within the coronary foreground and took the maximum across all cases as the global upper clipping value, which resulted in an upper bound of 1311 HU. Figure 3.13 illustrates, for a representative case, the effect of the three different upper limits.
After clipping to $[-330, 1311]$ HU, intensities were linearly mapped to the range $[-3, 3]$.

3.4.1.4 Binarization

All coronary artery vessel labels were merged into a single binary foreground class, and all remaining voxels were set to background. This also included the aorta, which was treated as background because it is not part of the coronary artery tree.

3.4.2 Network Training

All models are trained from scratch for 1000 epochs using 3D patches and evaluated using five-fold cross-validation on the training set. We train all networks with a generic baseline loss given by the sum of Dice loss and CE loss:

$$\mathscr{L}_{\text{generic}} = \mathscr{L}_{\text{Dice}} + \mathscr{L}_{\text{CE}}. \tag{3.12}$$

For the connectivity-preserving configurations, an additional connectivity term $\mathscr{L}_\ell$ is added to this baseline loss with a fixed weighting factor of 1,

$$\mathscr{L}_{\text{connectivity},\ell} = \mathscr{L}_{\text{generic}} + \mathscr{L}_\ell, \qquad \ell \in \{\text{SR}, \text{clDice}, \text{cbDice}, \text{clCE}\}, \tag{3.13}$$

where $\mathscr{L}_\ell$ denotes one of the connectivity-preserving losses described in Section 3.3.2.
In the exploratory pipeline, we compared the $\mathscr{L}_{\text{generic}}$ to the SR configuration $\mathscr{L}_{\text{connectivity,SR}}$, whereas the extended set of connectivity-preserving losses $\mathscr{L}_{\text{connectivity,SR}}$, $\mathscr{L}_{\text{connectivity,clDice}}$, $\mathscr{L}_{\text{connectivity,cbDice}}$ and $\mathscr{L}_{\text{connectivity,clCE}}$ was evaluated within the nnU-Net framework.
Given the small volumetric extent of the coronary arteries relative to the full CCTA volume, patch sampling is a crucial design choice, as it determines how often coronary voxels are seen during training and how much anatomical context is available within each patch. In the exploratory pipeline, patch centres are sampled based on precomputed bounding boxes around the SCCT segments (Figure 2.2), and an additional bounding box is defined for the entire volume. During training, a bounding box is first selected according to predefined sampling probabilities: 30% of patches are drawn from the bounding box of the entire volume, and the remaining 70% are distributed uniformly across all coronary segment boxes. Given the selected bounding box, the patch centre is then sampled uniformly within this box, and a fixed-size 3D patch of $72 \times 72 \times 72$ voxels is extracted around that centre. With this strategy, 78% of patches contained at least one coronary foreground voxel, while the average foreground proportion within a patch was 0.85%.
nnU-Net applies a built-in foreground oversampling mechanism. Given a batch size B and an oversampling fraction p, the number of foreground-enforced samples per batch is given by

$$\#\text{FG} = B - \text{round}\big(B(1-p)\big). \tag{3.14}$$

For our configuration with a patch size of $160 \times 160 \times 96$ voxels, a batch size of $B = 2$, and an oversampling fraction of $p = 0.33$, this setting resulted in 100% of patches containing at least one coronary foreground voxel, while the mean foreground voxel fraction per patch was 0.15%.
Table 3.5 summaries the key fixed training hyperparameters used for the final exploratory pipeline and the nnU-Net-based setup. Each augmentation is applied with a specified probability. Additional experiments with alternative architecture templates, optimizers and learning-rate schedules were conducted as part of preliminary ablation studies and are reported in Section 4.2.3.

Table 3.5: Comparison of key fixed training hyperparameters between the exploratory pipeline and the nnU-Net-based setup.

Hyperparameters	Exploratory pipeline	nnU-Net
Epochs	1000	1000
Epoch size	200	250
Batch size	8	2
Patch size	$72 \times 72 \times 72$	$160 \times 160 \times 96$
Normalization operation	Batch normalization	Instance normalization
Deep supervision	No	Yes
Data augmentation	Scaling, rotation, Gaussian noise, Gaussian blur, gamma	Scaling, rotation, Gaussian noise, Gaussian blur, gamma, contrast, low-resolution simulation, mirroring

3.4.3 Inference

During inference, prediction was performed patch-wise, consistent with the patch-based training strategy. Due to the limited receptive field, prediction quality typically decreases towards patch borders. To mitigate stitching artifacts, neighboring patches are overlapped. In this setup, the network is applied in a sliding-window fashion and processes windows of the same size as the patch size used during training. The impact of different overlap sizes on segmentation performance is quantitatively assessed in Section 4.1.
Predictions are merged by averaging the softmax outputs of the network across all patches covering a voxel. The aggregated probability for class c at voxel x is given by

$$p_{\text{final}}(c,x) = \frac{1}{K} \sum_{k=1}^{K} p_k(c,x), \tag{3.15}$$

where $p_k(c,x)$ denotes the softmax probability of class c from patch k, and K is the number of overlapping patches at voxel x. The final segmentation mask is obtained by a voxel-wise $\arg\max$ over $p_{\text{final}}(c,x)$.
nnU-Net additionally employs a Gaussian weighting scheme within the sliding-window inference that assigns higher weights to voxels near the patch center and lower weights near patch borders. Furthermore, it uses test-time augmentation by mirroring patches along all spatial axes.

3.4.4 Postprocessing

A connected component analysis (26-neighborhood connectivity) of the predicted foreground labels was performed to investigate different postprocessing strategies. In one variant, all connected components with fewer than 100 voxels are removed, treating these small clusters as noise artifacts. In an alternative variant, only the largest connected components are retained: typically the two largest components, corresponding to the left and right coronary

trees, while in the two cases with three coronary trees in the ground truth, all three largest components are preserved to match the underlying anatomy.
In contrast to the exploratory pipeline, nnU-Net is trained without spatial cropping during preprocessing. Empirically, we observed false-positive predictions near the borders of the original field-of-view when operating on full volumes. To reduce these artifacts, we subsequently apply the previously described heart-mask-based cropping (Section 3.4.1) as a postprocessing step for the nnU-Net segmentations, but without the additional 20-voxel margin. All cropping and postprocessing variants are quantitatively compared in Section 4.4.3.

3.5 Evaluation

To compare the performance of the different loss configurations within the proposed pipeline, we evaluate the models using quantitative scoring metrics and dedicated visualization techniques. The quality criteria comprise standard voxel-wise vessel mask accuracy as well as topology- and connectivity-aware measures that target vessel accuracy and centerline completeness. In addition, we report the runtime characteristics of each loss configuration. For qualitative assessment, we employ stretched Multiplanar Reformat (sMPR), which provides a standardized view along the entire vessel course and facilitates the interpretation of the quantitative metrics compared to conventional visualization views. Finally, a dedicated error analysis of FP and False Negative (FN) regions is conducted to characterize recurring failure patterns across the evaluated models.

3.5.1 Vessel Mask Accuracy

Vessel mask accuracy quantifies the voxel-wise agreement between the predicted coronary artery mask and the reference annotation.
As a volumetric overlap measure, we employ the Dice similarity coefficient. Let $\mathbf{y} \in \{0,1\}^n$ denote the binary reference mask and $\hat{\mathbf{p}} \in \{0,1\}^n$ the segmentation derived from the network output. Using the precision and recall notation introduced in Section 3.3.1.1, but evaluated on the binary masks $(\mathbf{y}, \hat{\mathbf{p}})$, we obtain

$$\pi(\mathbf{y}, \hat{\mathbf{p}}) = \frac{\mathbf{y}^\top \hat{\mathbf{p}}}{\hat{\mathbf{p}}^\top \mathbf{1}}, \qquad \rho(\mathbf{y}, \hat{\mathbf{p}}) = \frac{\mathbf{y}^\top \hat{\mathbf{p}}}{\mathbf{y}^\top \mathbf{1}}. \tag{3.16}$$

Here, $\mathbf{y}^\top \hat{\mathbf{p}}$ denotes the number of true-positive voxels, $\hat{\mathbf{p}}^\top \mathbf{1}$ the number of predicted foreground voxels, and $\mathbf{y}^\top \mathbf{1}$ the number of reference foreground voxels.
The Dice similarity coefficient can then be written as the harmonic mean of precision and recall,

$$\mathrm{Dice}(\mathbf{y}, \hat{\mathbf{p}}) = \frac{2}{\frac{1}{\pi(\mathbf{y}, \hat{\mathbf{p}})} + \frac{1}{\rho(\mathbf{y}, \hat{\mathbf{p}})}}. \tag{3.17}$$

To complement vessel mask accuracy, we additionally quantify spatial discrepancies using the Hausdorff Distance (HD) between surfaces derived from the binary masks $\mathbf{y}$ and $\hat{\mathbf{p}}$. The directed HD from the reference to the segmentation and vice versa is defined as

$$d(\mathbf{y},\hat{\mathbf{p}}) = \max_{x\in\mathbf{y}} \min_{z\in\hat{\mathbf{p}}} \|x-z\|_2, \qquad d(\hat{\mathbf{p}},\mathbf{y}) = \max_{z\in\hat{\mathbf{p}}} \min_{x\in\mathbf{y}} \|x-z\|_2. \tag{3.18}$$

The symmetric HD is then given by

$$\mathrm{HD}_{\mathrm{sym}} = \max\big(d(\mathbf{y},\hat{\mathbf{p}}), d(\hat{\mathbf{p}},\mathbf{y})\big). \tag{3.19}$$

In practice, several variants of the HD are commonly used in medical image segmentation. Among the most widely used are the 95th percentile HD (HD_{95}), the average surface distance ($\mathrm{HD}_{\mathrm{avg}}$), and one-sided directed variants. HD_{95} is defined as the 95th percentile of the bidirectional surface distance distribution. The average surface distance $\mathrm{HD}_{\mathrm{avg}}$ computes the mean boundary deviation between reference and prediction. A one-sided directed HD considers only the distances from the reference surface to the predicted surface ($\mathrm{HD}_{\mathrm{GT}}$). The use of these variants for evaluating coronary artery segmentations in CCTA is discussed in Section 4.2.2.

3.5.2 Vessel accuracy

Vessel accuracy assesses how well the predicted segmentation reproduces the course and branching pattern of the coronary artery tree, while tolerating small boundary deviations. To quantify vessel topology accuracy, we employ the clDice metric. Let $\mathbf{y} \in \{0,1\}^n$ denote the binary reference mask and $\hat{\mathbf{p}} \in \{0,1\}^n$ the segmentation derived from the network output. Conceptually, we reuse the same precision and recall structure as for the clDice loss in Section 3.3.2.1, but for evaluation purposes differentiability is no longer required, so we operate on binary masks and hard skeletons. The corresponding hard skeletons, $\mathbf{y}_s$ and $\hat{\mathbf{p}}_s$, are computed using the binary skeletonisation procedure described in Section 3.3.2. The associated precision and recall are defined as

$$\pi(\mathbf{y},\hat{\mathbf{p}}_s) = \frac{\hat{\mathbf{p}}_s^\top \mathbf{y}}{\hat{\mathbf{p}}_s^\top \mathbf{1}}, \qquad \rho(\mathbf{y}_s,\hat{\mathbf{p}}) = \frac{\mathbf{y}_s^\top \hat{\mathbf{p}}}{\mathbf{y}_s^\top \mathbf{1}}. \tag{3.20}$$

Here, $\pi(\mathbf{y},\hat{\mathbf{p}}_s)$ measures which fraction of the predicted skeleton lies inside the reference foreground (skeleton precision), whereas $\rho(\mathbf{y}_s,\hat{\mathbf{p}})$ measures which fraction of the reference skeleton is recovered by the prediction (skeleton recall). The clDice score is then given by their harmonic mean,

$$\mathrm{clDice}(\mathbf{y},\hat{\mathbf{p}}) = \frac{2}{\frac{1}{\pi(\mathbf{y},\hat{\mathbf{p}}_s)} + \frac{1}{\rho(\mathbf{y}_s,\hat{\mathbf{p}})}}. \tag{3.21}$$

To tolerate small boundary deviations of the vessel masks, we report an ε-relaxed Dice score, denoted by εDice. Let $\mathbf{y} \in \{0,1\}^n$ denote the binary reference mask and $\hat{\mathbf{p}} \in \{0,1\}^n$ the segmentation derived from the network output. We denote the sets of foreground voxels of reference and prediction by

$$R = \{x \mid \mathbf{y}(x) = 1\}, \qquad P = \{x \mid \hat{\mathbf{p}}(x) = 1\}. \tag{3.22}$$

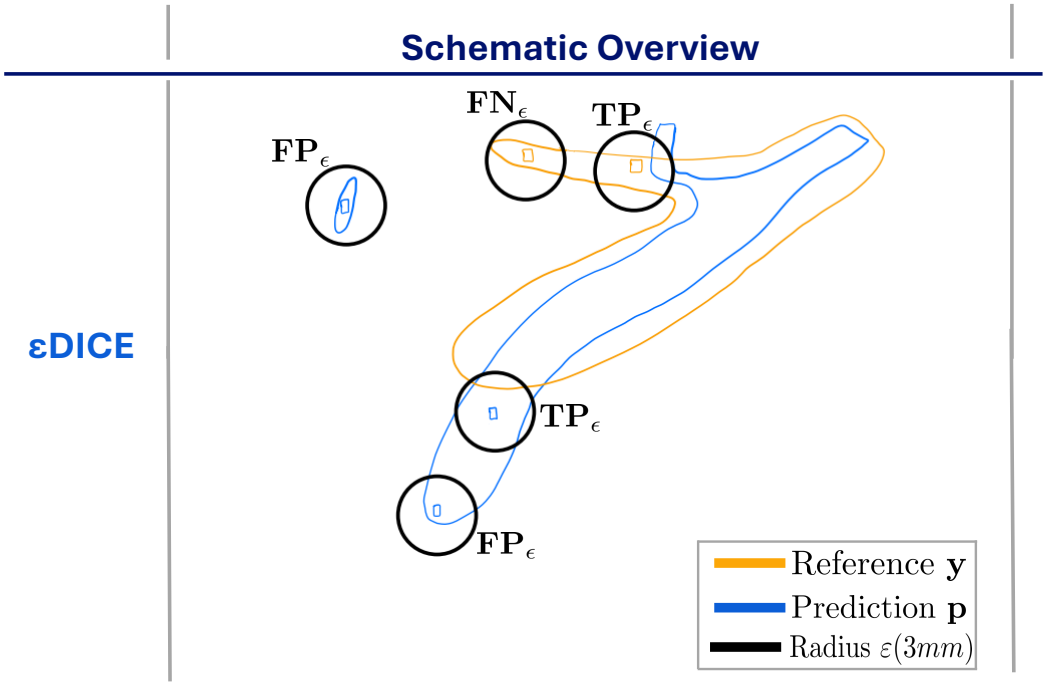

Figure 3.14: Schematic illustration of the ε-relaxed TP, FP and FN voxels for $\varepsilon = 3\,\text{mm}$. Predicted voxels with at least one reference voxel within 3 mm are counted as ε-TP, predicted voxels without such a neighbour become ε-FP, and reference voxels without any predicted voxel within 3 mm are counted as ε-FN.

We choose $\varepsilon = 3\,\text{mm}$, which roughly corresponds to the typical diameter of the major coronary arteries. For this fixed distance threshold, we define the ε-relaxed sets of TP, FP and FN as

$$\begin{aligned} \mathbf{TP}_\varepsilon &= \{\, p \in P \mid \exists r \in R : d(p,r) \le \varepsilon \,\}, \\ \mathbf{FP}_\varepsilon &= \{\, p \in P \mid \forall r \in R : d(p,r) > \varepsilon \,\}, \\ \mathbf{FN}_\varepsilon &= \{\, r \in R \mid \forall p \in P : d(r,p) > \varepsilon \,\}, \end{aligned} \tag{3.23}$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance between voxel centers. For improved readability, we use set notation for the ε-relaxed definitions. This is equivalent to representing the same voxel subsets by binary indicator vectors, as used throughout this work. For instance, intersections correspond to the Hadamard product of binary masks. A schematic overview of the ε-relaxed confusion sets is shown in Figure 3.14.
The corresponding ε-relaxed precision and recall are given by

$$\pi_\varepsilon(\mathbf{TP}_\varepsilon, \mathbf{FP}_\varepsilon) = \frac{|\mathbf{TP}_\varepsilon|}{|\mathbf{TP}_\varepsilon| + |\mathbf{FP}_\varepsilon|}, \qquad \rho_\varepsilon(\mathbf{TP}_\varepsilon, \mathbf{FN}_\varepsilon) = \frac{|\mathbf{TP}_\varepsilon|}{|\mathbf{TP}_\varepsilon| + |\mathbf{FN}_\varepsilon|}. \tag{3.24}$$

The εDice score is then defined by their harmonic mean,

$$\varepsilon\text{Dice} = \frac{2}{\dfrac{1}{\pi_\varepsilon(\mathbf{TP}_\varepsilon, \mathbf{FP}_\varepsilon)} + \dfrac{1}{\rho_\varepsilon(\mathbf{TP}_\varepsilon, \mathbf{FN}_\varepsilon)}}. \tag{3.25}$$

3.5.3 Centerline Completeness

Centerline completeness quantifies to which extent the predicted segmentation recovers the full extent and connectivity of the coronary artery centerline. It directly measures missing branches, disconnections and gaps along the vascular tree and therefore provides the most informative assessment of quality for coronary artery tree segmentation.
To quantify how much of the reference centerline is covered by the predicted vessel mask, we employ the centerline True Positive Rate (clTPR). The clTPR corresponds to the skeleton recall term of the clDice metric formulation (Section 3.5.2). Let $\mathbf{y}_s$ denote the binary reference skeleton and $\hat{\mathbf{p}}$ the binary segmentation derived from the network output. The clTPR is defined as

$$\mathrm{clTPR}(\mathbf{y}_s, \hat{\mathbf{p}}) = \frac{\mathbf{y}_s^\top \hat{\mathbf{p}}}{\mathbf{y}_s^\top \mathbf{1}}, \tag{3.26}$$

which measures the fraction of reference centerline voxels that are recovered by the predicted segmentation.
To specifically assess whether the correctly recovered part of the coronary tree contains disconnections, we compute the TP β_0 Error. Let $\mathbf{y} \in \{0,1\}^n$ denote the reference mask and $\hat{\mathbf{p}} \in \{0,1\}^n$ the predicted vessel mask, and define the TP mask as the voxel-wise intersection

$$\mathbf{TP} = \mathbf{y} \odot \hat{\mathbf{p}}\,. \tag{3.27}$$

Let $\beta_0(\cdot)$ denote the number of connected components of a binary mask under 26-neighborhood connectivity in 3D. The TP β_0 Error is then defined as

$$\text{TP-}\beta_0\text{E} = \left|\beta_0(\mathbf{TP}) - \beta_0(\mathbf{y})\right|. \tag{3.28}$$

To distinguish whether additional components correspond to small gaps that could be bridged by simple postprocessing or to major structural breaks, we compute the Gap Count (GC). Let $\{C_i\}$ denote the set of connected components of the true-positive vessel tree, and let $\mathscr{A}$ be the set of component pairs that are adjacent along the underlying coronary anatomy. For two components C_i and C_j we define their minimal distance as

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \|x - y\|_2\,. \tag{3.29}$$

To account for anisotropic voxel spacing, we use a spacing-aware distance threshold

$$\delta = \sqrt{s_x^2 + s_y^2 + s_z^2}, \tag{3.30}$$

where (s_x, s_y, s_z) denote the voxel spacing in mm and δ corresponds to the voxel diagonal length. The GC is then defined as

$$\mathrm{GC} = \sum_{(i,j) \in \mathscr{A}} [\![d(C_i, C_j) > 3\,\delta]\!], \tag{3.31}$$

where $[\![\cdot]\!]$ denotes the Iverson bracket. Only breaks with an inter-component distance larger than $3\,\delta$ are counted as gaps.

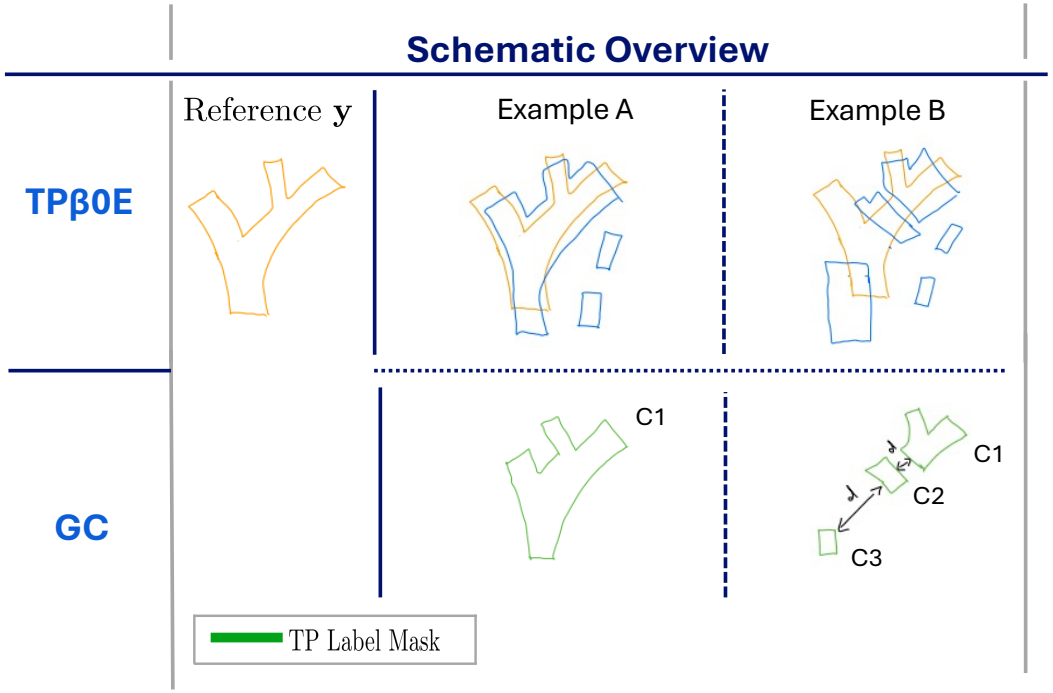

Figure 3.15: Schematic illustration of the TPβ_0 error and the GC. The upper row shows two examples where the true-positive vessel mask produces either no additional connected components (Example A) or two additional components (Example B), resulting in TPβ_0E = 0 and 2, respectively. The lower row illustrates the computation of the GC, where distances between anatomically adjacent connected components are evaluated to identify gaps between them.

A schematic illustration of the TPβ_0 error and the GC is provided in Figure 3.15.
Beyond the proposed centerline-based metrics, connectivity is often assessed in the literature using global topological descriptors such as Betti numbers [30]. The zeroth Betti number β_0 counts connected components, whereas the first Betti number β_1 counts loops ("holes") in a binary mask. Both can therefore be interpreted as global indicators of topological correctness and connectivity: deviations in β_0 reflect fragmented or spuriously merged components, while deviations in β_1 indicate the appearance or disappearance of loops.
Beyond the proposed centerline-based metrics, connectivity is often assessed in the literature using global topological descriptors such as Betti numbers. The zeroth Betti number β_0 counts connected components, whereas the first Betti number β_1 counts loops ("holes") in a binary mask. More recently, the Discrepancy between Intersection and Union (DIU) metric has been introduced in [78], which compares the topology of the union and intersection of prediction and reference and counts surplus as well as fragmented components that do not correspond one-to-one between the two shapes. In this sense, DIU aims to capture topological discrepancies in a more refined manner than simple Betti number errors. These additional topology-aware measures and their suitability for assessing coronary artery tree connectivity are discussed in Section 4.2.2.

3.5.4 Runtime

To assess the computational efficiency of the different loss configurations, we analyse their runtime behaviour.
For each model, we measure the total training time for 1000 epochs and the effective time-to-convergence, defined retrospectively as the time at which the validation Dice score no longer improves within the subsequent 100 epochs.

3.5.5 Stretched Multiplanar Reformats

For qualitative assessment of vessel continuity and overall segmentation quality, we employ sMPRs, which provide a straightened view of the coronary arteries and allow the reference and predicted vessel masks to be inspected jointly in a common layout.
The coronary artery tree is represented by a global centerline together with a segmentation of this centerline into individual anatomical vessels; for each vessel, a specific subsegment of the global centerline is considered, and along this subsegment local cross-sectional planes orthogonal to the centerline direction are defined. The underlying CCTA data are resampled on a regular grid in each plane, and the resulting cross-sections are stacked along the centerline parameter, yielding a 2D stretched representation in which one axis corresponds to vessel length and the other axis represents the cross-sectional extent of the lumen and wall. The same transformation is applied to the reference and predicted vessel masks, which enables a direct, vessel-wise comparison in a consistent layout. Compared to conventional 3D renderings or axial slices, this representation allows inspection of the entire vessel course from the proximal to the distal segments in a single, continuous view.

3.5.6 Error Analysis

To obtain a more fine-grained understanding of the failure modes, we perform a structured error analysis based on a categorization of FP and FN voxels.
For the FP analysis, we use the ε-relaxed confusion sets introduced in Section 3.5.2, thereby explicitly tolerating small boundary deviations between reference and prediction. In particular, we decompose the ε-false-positive set into three disjoint subclasses,

$$\mathbf{FP}_{\varepsilon} \;=\; \mathbf{FP}_{\text{prox}} \dot{\cup} \mathbf{FP}_{\text{dist}} \dot{\cup} \mathbf{FP}_{\text{float}}. \tag{3.32}$$

Let $C = \{c_i\}_{i=1}^{N}$ denote the centerline of the coronary tree together with its segmentation into individual vessels via the corresponding index ranges. Each voxel $p \in \mathbf{FP}_{\varepsilon}$ is mapped to the nearest centerline position

$$\nu(p) := \arg\min_{c \in C} d(p,c), \tag{3.33}$$

and the centerline points associated with root start locations (i.e. proximal origins of the major coronary vessels) and vessel termination points define the sets of proximal and distal

endpoints, $E_{\text{prox}} \subset C$ and $E_{\text{dist}} \subset C$, respectively. Using these sets, we obtain the following definitions:

$$\begin{aligned} \mathbf{FP}_{\text{prox}} &:= \big\{\, p \in \mathbf{FP}_\varepsilon \;\big|\; \nu(p) \in E_{\text{prox}} \,\big\}, \\ \mathbf{FP}_{\text{dist}} &:= \big\{\, p \in \mathbf{FP}_\varepsilon \;\big|\; \nu(p) \in E_{\text{dist}} \,\big\}, \\ \mathbf{FP}_{\text{float}} &:= \mathbf{FP}_\varepsilon \setminus \big(\mathbf{FP}_{\text{prox}} \cup \mathbf{FP}_{\text{dist}}\big). \end{aligned} \tag{3.34}$$

Here, $\mathbf{FP}_{\text{prox}}$ captures proximal oversegmentation near the coronary ostia, $\mathbf{FP}_{\text{dist}}$ corresponds to distal overextension beyond the annotated vessel ends, and $\mathbf{FP}_{\text{float}}$ comprises the remaining $\mathbf{FP}_\varepsilon$ voxels that appear as free-floating structures in the volume.

For the FN analysis, we deliberately set the distance threshold to $\varepsilon = 0$, which means that we operate on the original confusion sets. The rationale is that even small spatial tolerances would already bridge narrow gaps in the vessel mask and thereby hide false-negative voxels that cause connectivity breaks. In particular, we decompose the classical FN set into three disjoint subclasses,

$$\mathbf{FN} \;=\; \mathbf{FN}_{\text{disc}} \dot\cup \mathbf{FN}_{\text{thin}} \dot\cup \mathbf{FN}_{\text{short}}. \tag{3.35}$$

Let $\kappa_{26}(\cdot)$ denote the number of connected components of a binary mask under 26-neighborhood connectivity in 3D, and let $L_{\mathbf{FN}}(x)$ be the label of the connected component of $\mathbf{FN}$ that contains the voxel x. We denote the corresponding component mask by

$$\Omega_{L_{\mathbf{FN}}(x)} := \{\, y \mid L_{\mathbf{FN}}(y) = L_{\mathbf{FN}}(x) \,\}. \tag{3.36}$$

A false-negative voxel is classified as disconnectivity-inducing if its component reduces the number of connected components in the true-positive mask when hypothetically added back:

$$\mathbf{FN}_{\text{disc}} = \big\{\, x \in \mathbf{FN} \;\big|\; \kappa_{26}\big(\mathbf{TP} \cup \Omega_{L_{\mathbf{FN}}(x)}\big) \;<\; \kappa_{26}(\mathbf{TP}) \,\big\}. \tag{3.37}$$

The remaining false negatives are collected in

$$\mathbf{FN}_{\text{other}} := \mathbf{FN} \setminus \mathbf{FN}_{\text{disc}}. \tag{3.38}$$

To further distinguish locally thin from prematurely shortened vessels in the prediction relative to the reference, we use the skeleton of the reference. Let $Y_s := \text{supp}(\mathbf{y}_s) = \{x \in \Omega \mid \mathbf{y}_s(x) = 1\}$ denote the set of reference skeleton voxels. Using the TP indicator, we define

$$T_s := \{\, x \in Y_s \mid \mathbf{TP}(x) = 1 \,\}, \qquad F_s := Y_s \setminus T_s. \tag{3.39}$$

Each voxel $x \in \mathbf{FN}_{\text{other}}$ is projected onto the closest skeleton point via

$$\nu(x) := \arg\min_{s \in \mathbf{y}_s} d(x,s), \tag{3.40}$$

and we define

$$\begin{aligned} \mathbf{FN}_{\text{thin}} &= \{\, x \in \mathbf{FN}_{\text{other}} \;\mid\; \nu(x) \in T_{\text{s}} \,\}, \\ \mathbf{FN}_{\text{short}} &= \{\, x \in \mathbf{FN}_{\text{other}} \;\mid\; \nu(x) \in F_{\text{s}} \,\}. \end{aligned} \tag{3.41}$$

Intuitively, $\mathbf{FN}_{\text{disc}}$ captures gaps along reference vessels where missing voxels interrupt connected segments, $\mathbf{FN}_{\text{thin}}$ corresponds to regions where the prediction follows the reference vessel course but with a thinner lumen than the annotation, and $\mathbf{FN}_{\text{short}}$ corresponds to distal reference segments that are not reached by the prediction and therefore appear as prematurely truncated vessels.

3.6 Statistical Analysis

In this work, statistical analysis is used to compare measurements obtained from two models with different loss configurations evaluated on the same test cases. The resulting observations are therefore paired, and all inferential comparisons are performed on the paired per-case differences.

In paired designs, the most common parametric test is the paired t-test, which tests whether the mean of the paired differences equals zero and assumes that these differences are approximately normally distributed [79]. If this assumption is not reasonable, a common non-parametric alternative is the Wilcoxon signed-rank test, which tests whether the median paired difference equals zero and is more robust to non-normality and outliers[80].

To determine whether a parametric or non-parametric procedure is appropriate, all measurements obtained from the models with different loss configurations were analyzed. For each metric, both visual and formal normality assessments were performed. Specifically, Quantil-Quantil (Q-Q) plots were inspected and the Shapiro–Wilk test was applied to assess univariate normality. The Q-Q plots compare the empirical distribution of each measurement with a theoretical normal distribution. Data points lying approximately on the reference line indicate normality, whereas systematic deviations from this line suggest non-normality. In line with standard practice, a p-value greater than 0.05 in the Shapiro–Wilk test is interpreted as indicating that the data are consistent with a normal distribution, whereas a p-value smaller than 0.05 indicates a significant deviation from normality [81].

Beyond the choice of parametric versus non-parametric tests, hypotheses can be formulated either as superiority or non-inferiority. In a superiority test, the null hypothesis states that there is no improvement, and we seek evidence that one loss configuration performs better than the other. In a non-inferiority test, the goal is to show that a loss configuration is not worse than a reference loss configuration by more than a predefined margin $\Delta > 0$. The margin Δ must be chosen a priori based on practical relevance.

We assume that all measurements in our dataset represent independent observations. Therefore, no tests accounting for data structure were required.

Chapter 4

Results and Interpretation

This chapter presents the experimental results and their interpretation in a stepwise progression toward testing the central research hypothesis. Before conducting the final statistical comparison, we first perform a series of preparatory experiments to establish a robust evaluation pipeline and to reduce confounding influences on the loss comparison. Section 4.1 analyzes the impact of sliding-window patch overlap on connectivity and inference time and motivates the overlap setting used thereafter. Section 4.2 reports findings from the exploratory pipeline, including an assessment of metric distributions and suitability, and derives stable blueprint choices for the nnU-Net framework. Building on this controlled setup, Section 4.3 provides a systematic comparison of connectivity-preserving loss formulations under a unified training configuration. Section 4.4 investigates the sensitivity of the selected loss to design choices such as input encoding, recall targets, and postprocessing. Finally, Section 4.5 conducts the statistical comparison between $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ to test the central research hypothesis.

4.1 Patch Overlap Study

Inference follows the patch-wise sliding-window scheme described in Section 3.4.3. In this setting, the volume is reassembled from locally predicted windows, which can introduce stitching artifacts at patch borders and, in turn, lead to connectivity breaks in the coronary artery tree after discretization. To study how strongly the overlap parameter controls these patch-induced effects, we evaluated three overlap configurations (0%, 25%, and 50%) and analyzed their impact on the preservation of coronary artery tree connectivity.

To quantify connectivity preservation under different overlap settings, we measured the TPβ_0 error. Since larger overlaps reduce the effective stride between patches, they increase the number of windows processed during sliding-window inference and thus the overall inference time. We therefore additionally compared the per-volume inference times to assess the computational impact of each configuration. Figure 4.1 shows the comparison between the patch-overlap settings with respect to the TPβ_0 error and the per-volume inference time.

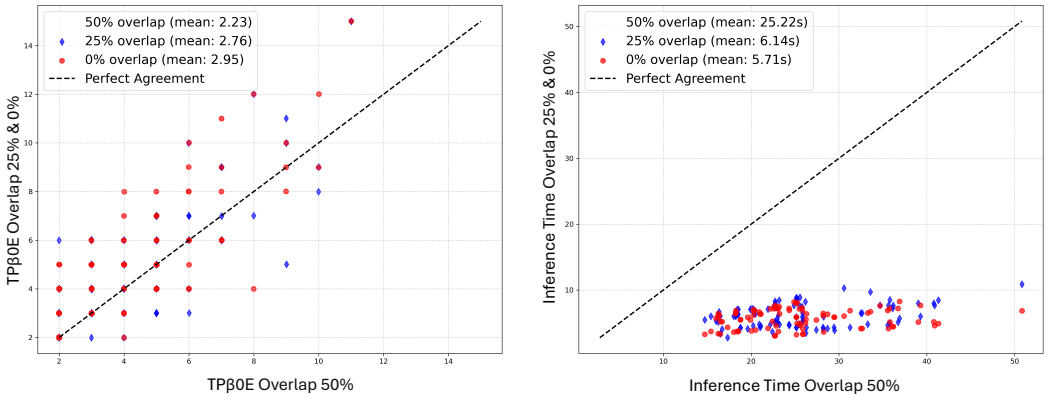

Figure 4.1: Scatter plots comparing the patch-overlap settings. Each point corresponds to one case. Blue diamonds show the values for 25% overlap plotted against the corresponding 50% values, and red circles show the values for 0% overlap plotted against the 50% values. The left panel shows the TPβ_0 error, whereas the right panel shows the corresponding per-volume inference time. Points on the diagonal indicate identical values for the compared settings. Points above the diagonal indicate higher errors or longer runtimes for 0% or 25% overlap than for 50% overlap, and points below the diagonal indicate lower values than the 50% reference.

The scatter plots show a consistent effect of the overlap configuration on connectivity and runtime. For the TPβ_0 error, most points lie above the diagonal when comparing 0% and 25% overlap with the 50% configuration. This indicates that 0% and 25% overlap tend to yield higher TPβ_0 errors than 50% overlap, which is also reflected in the mean values of 2.95, 2.76, and 2.23 for 0%, 25%, and 50% overlap, respectively. In contrast, the runtime comparison shows that 0% and 25% overlap yield almost identical inference times, with mean values of 5.71 s and 6.14 s, respectively, while 50% overlap results in substantially higher runtimes, with a mean inference time of 25.22 s.

These observations support that larger overlaps mitigate border-induced stitching artifacts and thereby reduce connectivity breaks after discretization. However, increasing the overlap from 25% to 50% substantially increases inference time (mean 6.14 s vs. 25.22 s). In this setup, inference accounts for only $\approx 1.5\%$ of the end-to-end runtime (preprocessing + training + inference + evaluation). Therefore, this overhead is negligible. Since connectivity preservation is the primary objective of this work, we use 50% overlap in all subsequent experiments.

4.2 Baseline Experiments

To establish a reference point for the subsequent large-scale experiments, we first performed a set of baseline studies within the exploratory pipeline using the setup described in Section 3.4. These experiments provide an initial comparison between the generic loss and a connectivity-preserving SR configuration (as defined in Section 3.4.2), allow us to characterize the distributional properties of the employed evaluation metrics and assess their suitability.

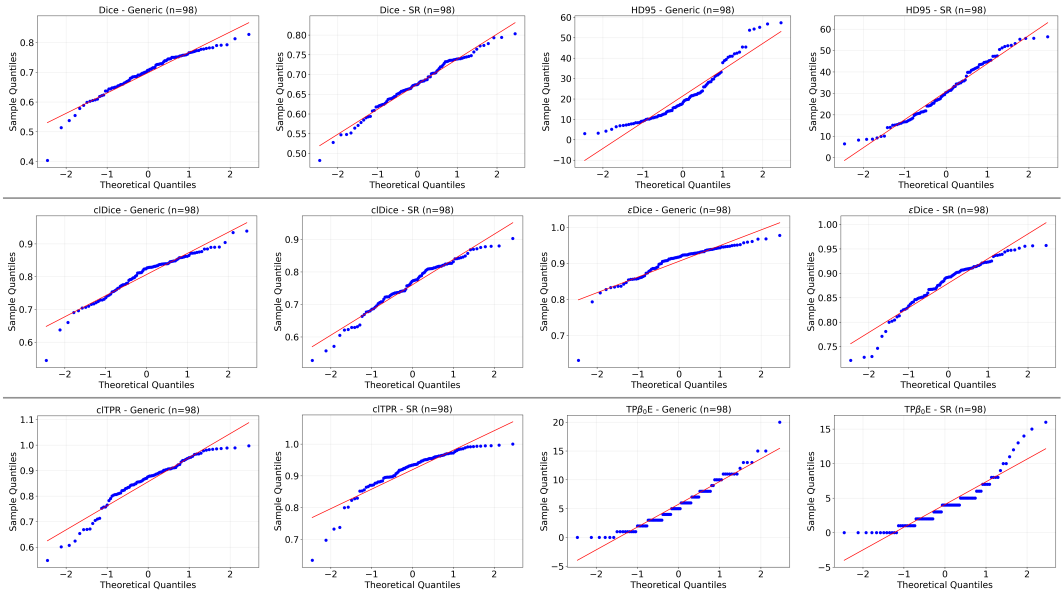

Figure 4.2: Q-Q plots of the case-wise distributions of selected evaluation metrics. The top row shows Dice and HD_{95} as representative vessel mask accuracy metrics, the middle row shows clDice and εDice as representative vessel accuracy metrics, and the bottom row shows clTPR and $TP\beta_0E$ as representative centerline completeness metrics. For each metric, the left plot corresponds to $\mathscr{L}_{\text{generic}}$ (labelled *Generic*) and the right plot to $\mathscr{L}_{\text{connectivity},SR}$ (labelled *SR*).

Finally, they support the selection of a stable blueprint configuration that will be reused in the later nnU-Net experiments.

4.2.1 Metric Distributions

To determine how the results should be reported, we first assessed the distribution of the evaluation metrics. For each metric and for both loss configurations, $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$, we visually inspected Q-Q plots and applied the Shapiro–Wilk test for normality. Representative Q-Q plots are shown in Figure 4.2, and the corresponding p-values are reported in Table 4.1.

Apart from the Dice score and HD_{95} under $\mathscr{L}_{\text{connectivity},SR}$, all metrics exhibited significant deviations from normality. The remaining evaluation metrics were assessed analogously and likewise did not satisfy the normality assumption. To ensure a consistent presentation across all metrics, we therefore report all results as median and interquartile range Q1–Q3.

4.2.2 Metric suitability

We next examined which evaluation metrics are suitable for comparing the segmentation performance obtained with a generic and a connectivity-preserving loss. To this end, we first

Table 4.1: Shapiro–Wilk normality test for each metric and loss configuration. Reported are the p-values for the null hypothesis of normality. Bold values indicate significant deviation from normality ($p < 0.05$).

Metric	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	$\mathbf{2.40 \times 10^{-4}}$	2.66×10^{-1}
εDice	$\mathbf{2.48 \times 10^{-9}}$	$\mathbf{4.45 \times 10^{-5}}$
clDice	$\mathbf{5.58 \times 10^{-4}}$	$\mathbf{9.30 \times 10^{-3}}$
clTPR	$\mathbf{2.69 \times 10^{-5}}$	$\mathbf{2.82 \times 10^{-8}}$
HD_{95}	$\mathbf{3.81 \times 10^{-6}}$	5.19×10^{-2}
$TP\beta_0E$	$\mathbf{7.03 \times 10^{-4}}$	$\mathbf{2.78 \times 10^{-7}}$

Metric	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	70.68 [65.90, 75.10]	67.47 [63.89, 72.70]
HD_{95}	18.15 [11.17, 28.28]	30.43 [20.62, 41.24]
clDice	82.55 [76.68, 84.78]	77.21 [72.28, 81.77]
εDice	91.80 [88.47, 93.68]	89.21 [85.10, 91.36]
clTPR	87.59 [81.47, 91.18]	93.38 [89.51, 96.24]
β_0E	24 [19, 31]	33 [26, 48]
β_1E	0 [0, 0]	0 [0, 0]
$TP\beta_0E$	5 [3, 8]	4 [2, 5]
DIU	24 [19, 31]	33 [27, 49]

Table 4.2: Median and interquartile range Q1–Q3 of all evaluation metrics applied to $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,SR}}$. The first block contains vessel mask accuracy metrics, the second block contains vessel accuracy metrics, and the third block contains centerline completeness metrics.

applied all evaluation metrics introduced in Section 3.5 to both $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$. The resulting median values and interquartile ranges (Q1–Q3) are summarized in Table 4.2.

For the vessel mask accuracy metrics, the Dice score is slightly higher under $\mathscr{L}_{\text{generic}}$ than under $\mathscr{L}_{\text{connectivity},SR}$ (Table 4.2). In contrast, HD_{95} shows a pronounced systematic increase for $\mathscr{L}_{\text{connectivity},SR}$, rising from 18.15 mm to 30.43 mm, with comparatively large values for both configurations. To clarify this behavior, Figure 4.3 shows the case with the largest HD_{95} for each loss configuration.

Across this and many other cases, HD_{95} is dominated by peripheral FP structures located at the image borders. These components form large distant clusters whose maximal surface distance to the reference substantially inflates HD_{95}. Although HD_{95} is intended as a robust variant of the HD, in segmentation of thin structures a small number of border FPs can dominate the 95th percentile and thereby obscure otherwise reasonable vessel delineation in the Region of Interest (ROI). Under $\mathscr{L}_{\text{connectivity},SR}$, these peripheral FPs occur more frequently and tend to be larger, which aligns with the recall-type nature of the added SR term that does not penalize FPs. HD_{avg} is far less influenced by such outliers, and HD_{GT} is

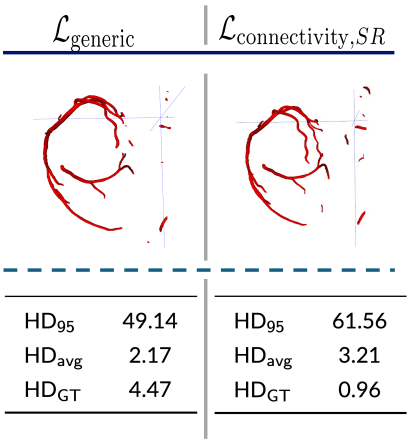

Figure 4.3: Representative case with the largest HD_{95} for $\mathscr{L}_{\text{generic}}$ (left) and $\mathscr{L}_{\text{connectivity},SR}$ (right). Top: Volume rendering of the predicted coronary artery tree. Bottom: Corresponding values for HD_{95}, HD_{avg}, and HD_{GT}.

not affected by boundary FPs at all. As seen in the example case, both decrease substantially for both predictions. These variants are less sensitive to boundary FPs by construction, but were not analyzed further in this work. We therefore use the Dice score as the representative vessel mask accuracy metric, i.e. as our primary measure of volumetric overlap.

The vessel accuracy metrics clDice and εDice show a similar pattern for both loss configurations. Both attain higher values than the standard Dice score, indicating that the vascular structures and their topology are largely preserved in the segmentations. While these metrics are still influenced by spurious structures at the image borders, they explicitly tolerate small boundary shifts as well as mild under- and over-segmentation, making them less sensitive to minor local discrepancies than purely mask-based measures.

clTPR exhibits the strongest difference between the two loss configurations, with clearly higher values under $\mathscr{L}_{\text{connectivity},SR}$ than under $\mathscr{L}_{\text{generic}}$. This indicates that the connectivity-preserving loss substantially improves the recovery of the ground-truth centerline. In the literature, the β_0E is commonly used as a connectivity metric. In our setting, however, β_0E partly reflects the boundary FPs discussed above rather than connectivity of the coronary tree itself. To mitigate this, we consider the TPβ_0E, which restricts the analysis to connected components that overlap the ground-truth tree. For both loss configurations, the first Betti error β_1E is zero, indicating that loops do not occur in our masks and that this metric can be neglected for the comparison. The DIU metric was originally proposed to provide a more refined topology-aware assessment than simple β_0-based errors. While the original work demonstrated examples where DIU captured discrepancies that β_0E missed, our results in Table 4.2 show very similar median values and interquartile ranges for DIU and β_0E across both loss configurations, indicating that DIU does not provide substantial additional discriminative power in our setting. To keep the metric set focused on clinically

Metric	U-Net		F-Net	
	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	70.68 [65.90, 75.10]	67.47 [63.89, 72.70]	69.47 [64.73, 73.40]	63.74 [56.98, 69.26]
clDice	82.55 [76.68, 84.78]	77.21 [72.28, 81.77]	78.33 [72.71, 82.31]	70.51 [61.16, 77.30]
εDice	91.80 [88.47, 93.68]	89.21 [85.10, 91.36]	91.58 [88.89, 93.71]	86.12 [79.45, 90.28]
clTPR	87.59 [81.47, 91.18]	93.38 [89.51, 96.24]	87.52 [80.78, 92.16]	94.76 [91.13, 97.52]
TPβ_0E	5 [3, 8]	4 [2, 5]	8 [4, 12]	4 [1, 7]

Table 4.3: Comparison of the U-Net and F-Net architecture templates under the generic loss $\mathscr{L}_{\text{generic}}$ and the connectivity-preserving loss $\mathscr{L}_{\text{connectivity,SR}}$. Values denote median [Q1, Q3] over all cases.

relevant connectivity properties, we therefore use clTPR and TPβ_0E as the primary centerline completeness measures in the subsequent analyses.

4.2.3 Blueprint Parameters

The nnU-Net framework defines a set of blueprint parameters, which are fixed architectural and training presets that are not automatically adapted to a given dataset. One of these fixed parameters is the architecture template, which specifies the overall network family from which the final model is instantiated. To select an appropriate template, we compared the standard U-Net architecture with an F-Net architecture. An F-Net consists of multiple resolution levels, where each level is defined by an encoder and a decoder. The outputs of the encoders are integrated in a dedicated feature-integration pathway, which is implemented as a series of decoders defined by the respective decoder field of each network level. To select an appropriate architecture, we compared the U-Net with the F-Net template under both $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,}SR}$. Table 4.3 summarizes the median performance for both templates and both loss configurations.

Since clTPR is the only metric that improves under the F-Net template when using the connectivity-preserving loss, and the improvement is limited to approximately 1.4 percentage points, we select the U-Net template as the architectural blueprint.

Another blueprint parameter concerns the choice of optimizer. While the exploratory pipeline employed Adadelta, widely used alternatives in medical image segmentation include AdamW and SGD with Nesterov momentum [82]. To determine an appropriate optimizer for the subsequent experiments, we compared these three methods. To keep the comparison focused and concise, we report only the median values under the connectivity-preserving loss $\mathscr{L}_{\text{connectivity,}SR}$ and use one representative metric for each category (Dice for vessel mask accuracy, εDice for vessel accuracy, and TPβ_0E for centerline completeness).

AdamW performs worst across all three metrics and is therefore not considered further. Compared to SGD with Nesterov momentum, Adadelta yields slightly higher Dice and

Metric	Adadelta	AdamW	SGD (Nesterov)
Dice	67.47	63.61	66.65
εDice	89.21	85.55	88.15
TPβ_0E	4	5	3

Table 4.4: Median performance of different optimizers under the connectivity-preserving loss $\mathscr{L}_{\text{connectivity,SR}}$ for representative metrics.

Metric	PolyLR	CosineAnnealingLR
Dice	66.65	66.86
εDice	88.15	88.25
TPβ_0E	3	4

Table 4.5: Comparison of PolyLR and cosine annealing learning rate schedulers under SGD with Nesterov momentum and $\mathscr{L}_{\text{connectivity,SR}}$. Values denote medians of representative metrics over all cases.

εDice scores (by roughly 1 percentage point), but at the cost of an increased TPβ_0E (4 vs. 3), indicating more connectivity-related component errors. Given the limited magnitude of these differences and the exploratory nature of this comparison, the results should be interpreted with caution. However, since nnU-Net by default employs SGD with Nesterov momentum and this optimizer achieves the lowest TPβ_0E while maintaining competitive overlap scores, we adopt SGD with Nesterov momentum as the optimizer for the nnU-Net-based framework in all subsequent experiments.

The last blueprint parameter we investigated is the learning rate scheduling strategy. In the experiments above, the PolyLR scheduler with an initial learning rate of 0.01 was employed together with the SGD optimizer. Alternative scheduling schemes, most notably cosine annealing, are frequently used in conjunction with SGD due to their smoother decay behavior and their tendency to stabilize late-stage optimization [83]. To assess whether cosine annealing provides an advantage for our task, we compared PolyLR and CosineAnnealingLR under the SGD optimizer with Nesterov momentum and the connectivity-preserving loss $\mathscr{L}_{\text{connectivity,}SR}$. As in the optimizer comparison, we report one representative metric per category. The resulting median values are summarized in Table 4.5.

Since training with the cosine annealing scheduler does not yield a relevant improvement in overlap metrics and even increases the TPβ_0E, we retain PolyLR as the learning rate scheduler for all subsequent experiments.

Loss Configuration	Dice	clDice	εDice
$\mathscr{L}_{\text{generic}}$	**78.76** [73.32, 81.42]	87.32 [83.92, 90.26]	93.58 [91.53, 95.71]
$\mathscr{L}_{\text{connectivity,SR}}$	78.31 [74.86, 81.25]	86.95 [82.81, 89.34]	93.61 [91.18, 95.38]
$\mathscr{L}_{\text{connectivity,clDice}}$	78.40 [73.48, 80.88]	**87.87** [84.37, 90.32]	93.64 [91.27, 95.81]
$\mathscr{L}_{\text{connectivity,cbDice}}$	78.06 [73.12, 81.44]	87.11 [83.58, 90.24]	**94.08** [91.15, 95.69]
$\mathscr{L}_{\text{connectivity,clCE}}$	78.28 [73.51, 81.46]	87.61 [83.57, 89.93]	93.57 [91.46, 95.89]

Table 4.6: Overlap-based metrics (Dice, clDice, εDice) for the generic loss and its extensions with connectivity-preserving terms. Values denote medians with interquartile ranges (Q1, Q3). Best median per metric is highlighted in bold.

4.3 Connectivity-Preserving Loss Functions

In contrast to the exploratory pipeline, which compared only the generic loss and the SR configuration, we here employ the nnU-Net framework to systematically assess the complete set of connectivity-aware losses introduced in Section 3.3.2. We use the blueprint parameters established in Section 4.2.3 and let nnU-Net determine the remaining hyperparameters based on the data fingerprint, as summarized in Table 3.5. As explained in Section 3.4.1.1, the built-in cropping procedure of nnU-Net has no effect on our dataset. To accelerate training, we therefore apply our custom heart-mask-based cropping strategy, resulting in an effective patch size of $160 \times 128 \times 112$ voxels. Apart from this modification, the overall training and inference workflow follows the setup described in Section 3.4. In all configurations, the generic loss $\mathscr{L}_{\text{generic}}$ serves as the baseline, and the respective connectivity-preserving term is added in an unweighted manner, as defined in Equation 3.13.

Table 4.6 and the boxplots in Fig. 4.4 show that the volumetric overlap metrics, which quantify vessel mask accuracy (Dice) and vessel accuracy (εDice, clDice), remain largely invariant under the choice of connectivity-preserving loss. Median values differ by less than one percentage point across all configurations and metrics, and the distributions show substantial overlap. These observations indicate that the connectivity-preserving terms do not adversely affect volumetric accuracy. The remaining differences lie within the natural variability of stochastic training.

A markedly different behavior is observed for the connectivity metrics, summarized in Table 4.7 and visualized in Fig. 4.5. The $\mathscr{L}_{\text{connectivity},SR}$ achieves the highest clTPR values with a median of 93.0 %, shifting the entire distribution upward and reducing the number of low-performing outliers. At the same time, $\mathscr{L}_{\text{connectivity},SR}$ yields the lowest TPβ_0E values with a median of 1.0, indicating fewer fragmented or spurious components. The $\mathscr{L}_{\text{connectivity},clDice}$ term provides only a modest improvement over the generic loss, increasing clTPR from 89.3 % to 90.2 %, but does not reduce the TPβ_0E median. In contrast, the $\mathscr{L}_{\text{connectivity},cbDice}$ term exhibits the weakest connectivity performance among all configurations, with a median

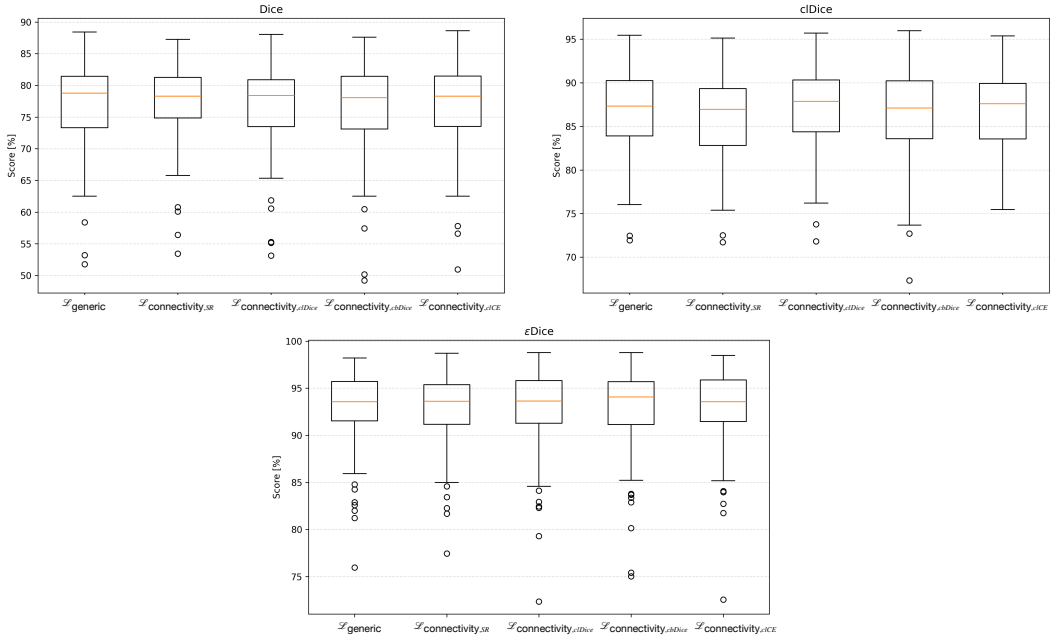

Figure 4.4: Boxplots of Dice, clDice, and εDice for the generic loss and all connectivity-preserving losses.

Loss Configuration	clTPR	TPβ_0E
$\mathcal{L}_{\text{generic}}$	89.32 [83.88, 93.50]	2 [1, 3]
$\mathcal{L}_{\text{connectivity,SR}}$	**93.01** [88.21, 96.84]	**1** [0, 3]
$\mathcal{L}_{\text{connectivity,clDice}}$	90.19 [84.69, 94.39]	2 [0, 3]
$\mathcal{L}_{\text{connectivity,cbDice}}$	86.56 [80.77, 90.64]	3 [1, 4]
$\mathcal{L}_{\text{connectivity,clCE}}$	89.26 [83.04, 93.41]	2 [1, 4]

Table 4.7: Connectivity metrics (clTPR and TPβ_0E) for the generic loss and its extensions with connectivity-preserving terms. Values denote medians with interquartile ranges (Q1, Q3). Best median per metric is highlighted in bold.

clTPR of 86.6 % and a TPβ_0E of 3. Notably, both $\mathcal{L}_{\text{connectivity},cbDice}$ and $\mathcal{L}_{\text{connectivity},clCE}$ perform even worse than the generic loss in this setup.

To complement the quantitative connectivity metrics, Fig. 4.6 shows sMPRs of the major vessels RCA and LCx for the same case, with overlaid reference and predicted segmentations for all loss configurations. For the RCA, only the $\mathcal{L}_{\text{connectivity},SR}$ model yields a fully continuous segmentation of both the main vessel and the small marginal side branch. All other configurations exhibit a connectivity break in the marginal branch and at least one interruption in the main vessel, with $\mathcal{L}_{\text{connectivity},cbDice}$ even showing two interruptions. Moreover, the gap produced by the generic loss configuration is markedly larger than those observed for the connectivity-preserving losses. For the latter, the discontinuities typically

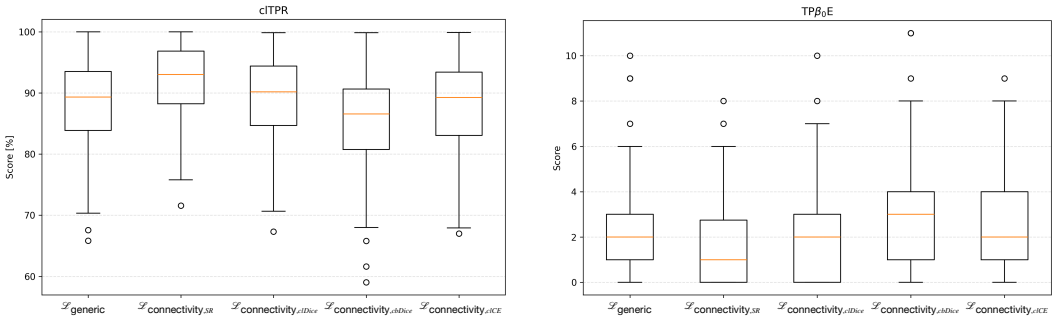

Figure 4.5: Boxplots of clTPR and TPβ_0E for the generic loss and all connectivity-preserving losses.

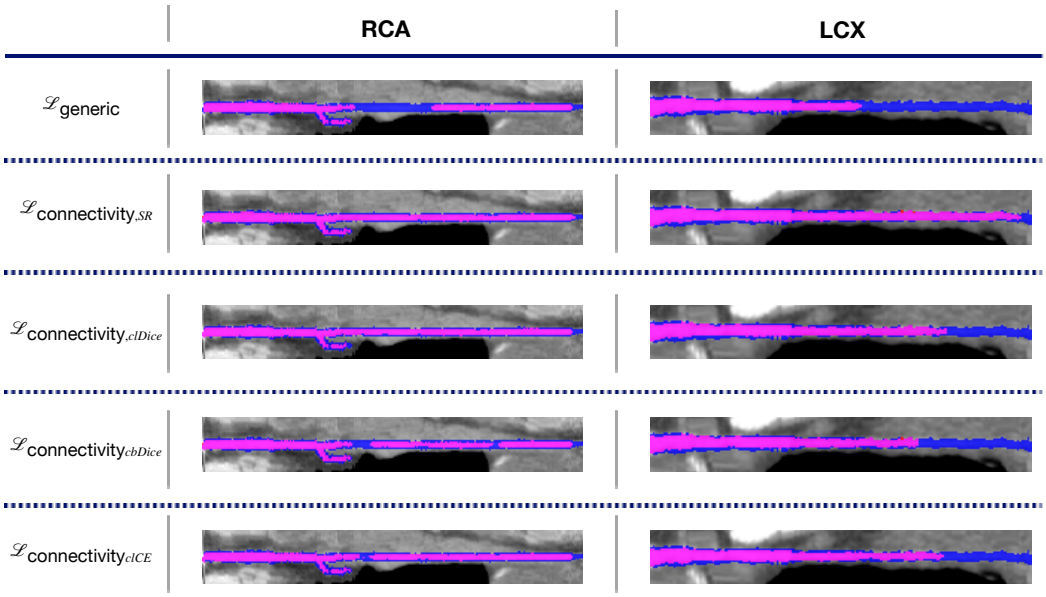

Figure 4.6: sMPRs of the RCA and LCx for the same test case, with reference segmentations in blue and voxels where reference and prediction overlap appear in magenta.

correspond to short gaps that could, in principle, be closed by light-weight postprocessing, whereas the generic loss often leaves extended missing segments. This distinction between small, potentially recoverable gaps and substantial discontinuities motivated the introduction of the GC metric (see Section 3.5.3), which is subsequently employed as an additional connectivity descriptor in the final experiment in Section 4.5.

Beyond improved connectivity, we also observe that $\mathscr{L}_{\text{connectivity},SR}$ consistently produces vessel predictions that follow the distal course more completely and do not terminate prematurely. This effect is clearly visible in the LCx example, where several alternative loss configurations stop too early, while the SR-based model adheres more closely to the reference.

To directly contrast the worst performing connectivity-preserving loss with the best performing configuration, we consider the case in which $\mathscr{L}_{\text{connectivity},cbDice}$ attains its lowest clTPR

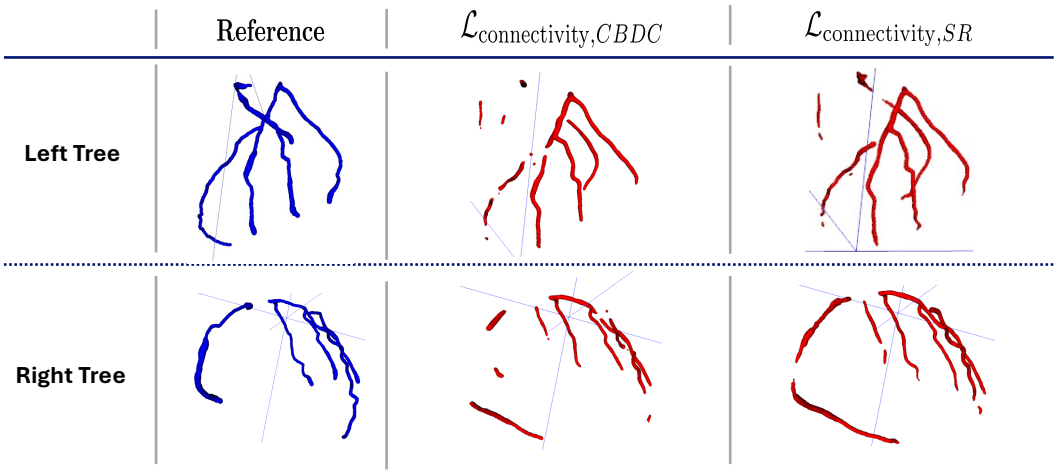

Figure 4.7: Qualitative comparison of the worst-performing $\mathscr{L}_{\text{connectivity},cbDice}$ case with the corresponding $\mathscr{L}_{\text{connectivity},SR}$ prediction. The reference is shown in blue, the predictions in red. All results are visualised as 3D-rendered volumes. Column 1 shows the reference, column 2 the prediction obtained with the $\mathscr{L}_{\text{connectivity},cbDice}$, and column 3 the prediction obtained with the $\mathscr{L}_{\text{connectivity},SR}$. The top row visualises the left coronary tree view, the bottom row the right coronary tree view.

and highest TPβ_0E. Fig. 4.7 compares the prediction obtained with $\mathscr{L}_{\text{connectivity},cbDice}$ for this case to the corresponding $\mathscr{L}_{\text{connectivity},SR}$ prediction.
For the left coronary tree, the cbDice-based model fails to preserve continuity at several locations: side branches are fragmented or terminate prematurely, and even the main RCA trunk is interrupted before reaching its distal course. In contrast, the SR-based prediction yields a much more continuous vessel tree, with extended distal segments and largely preserved branch connectivity. The degradation becomes even more pronounced in the right coronary tree, where the cbDice model produces anatomically implausible structures that no longer resemble coronary vessels. The SR-based prediction, on the other hand, reproduces the overall right coronary geometry, exhibiting only a single residual discontinuity.

It is important to note that the original formulations of clDice [20], cbDice [33], and clCE [32] differ from the uniform, unweighted setup used in this work. In their respective publications, these loss terms are combined with weighting schemes that control the relative strength of the generic and connectivity-preserving components. In our notation, such configurations can be expressed as

$$\mathscr{L}_\ell = \alpha\,\mathscr{L}_{\text{CE}} + \beta\,\mathscr{L}_{\text{Dice}} + \lambda\,\mathscr{L}_{\text{conn},\ell}, \qquad \ell \in \{\text{clDice}, \text{cbDice}, \text{clCE}\}, \tag{4.1}$$

with method-specific choices of α, β, and λ as proposed in the original papers.
For clDice and clCE, the authors set $\alpha = 0$, $\beta = 1$, and $\lambda = 1$, so that the generic loss does not include a CE term and is combined with a unit-weighted connectivity term. In contrast, cbDice includes both CE and Dice in the generic loss. For 3D binary segmentation problems, the authors propose either $(\alpha, \beta, \lambda) = \left(1, \frac{1}{2}, \frac{1}{2}\right)$, so that the generic loss is dominated by the CE term while the Dice and connectivity components receive equal

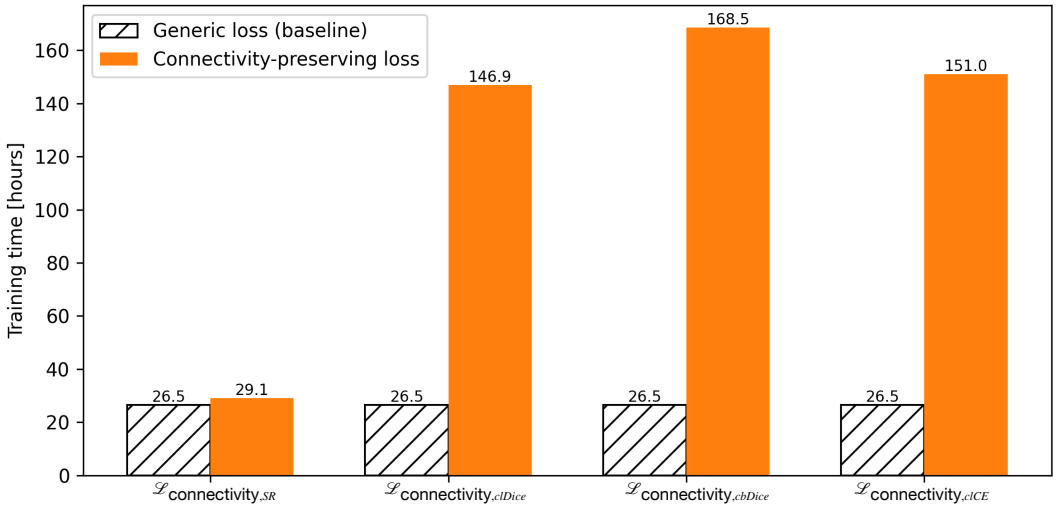

Figure 4.8: Comparison of the total training time for the fastest fold of each loss configuration. The hatched bars indicate the runtime of the $\mathscr{L}_{\text{generic}}$ baseline, while the solid bars show the corresponding runtime when adding a connectivity-preserving term.

weight, or $(\alpha, \beta, \lambda) = \left(1, \frac{1}{3}, \frac{2}{3}\right)$, which further down-weights the Dice term and emphasizes the connectivity component relative to Dice, while CE still carries the largest weight.
It is possible that using their original settings would lead to improved performance compared to the results obtained with the unified configuration used in this study. However, the unified weighting ensures that the observed differences primarily reflect the specific behavior of the loss terms rather than differences in hyperparameter choices.

Beyond segmentation accuracy and connectivity, the computational cost of each loss configuration must also be taken into account. In this context, the training time is largely insensitive to the specific choice of weighting parameters, as these only rescale existing loss terms but do not change the underlying operations. To this end, we assessed the training runtime of the fastest fold for all losses, using the generic loss as the baseline, as shown in Figure 4.8. $\mathscr{L}_{\text{connectivity},SR}$ increases training time only marginally, from 26.5 h to 29.1 h, whereas $\mathscr{L}_{\text{connectivity},clDice}$ and $\mathscr{L}_{\text{connectivity},clCE}$ require roughly six times and $\mathscr{L}_{\text{connectivity},cbDice}$ even about seven times the baseline runtime. This behavior is consistent with the underlying implementations of the connectivity terms. For $\mathscr{L}_{\text{connectivity},SR}$, the hard reference skeleton is precomputed on the CPU during data loading, such that the loss evaluation during training only involves an additional mask-based overlap computation. In contrast, $\mathscr{L}_{\text{connectivity},clDice}$, $\mathscr{L}_{\text{connectivity},clCE}$, and $\mathscr{L}_{\text{connectivity},cbDice}$ rely on a differentiable soft-skeletonisation operator applied on the GPU to both the reference and the prediction in every training iteration, which substantially increases the computational burden. The runtime of $\mathscr{L}_{\text{connectivity},cbDice}$ is highest because, in addition to the skeletonization, a three-dimensional Euclidean distance transform has to be computed. In our implementation, we adopt the soft skeletonisation scheme proposed in [34], which provides high topological fidelity at the expense of runtime. Using a

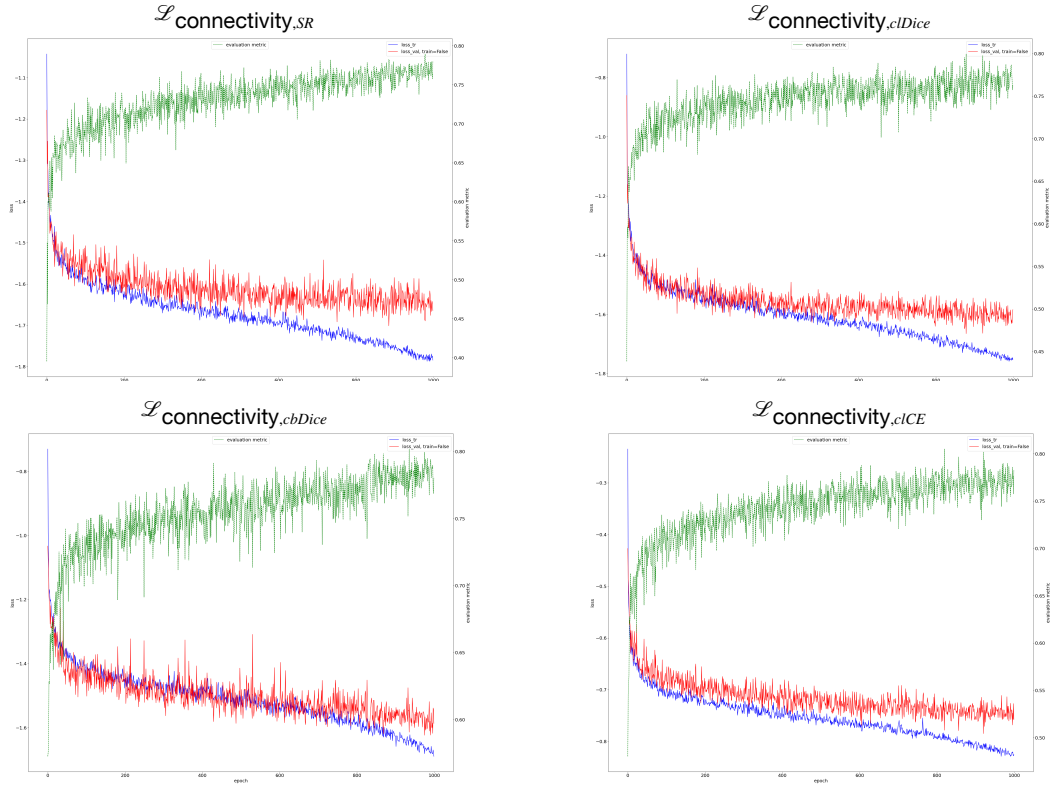

Figure 4.9: Training and validation curves for the nnU-Net models with the four connectivity-preserving loss configurations. The panels show, from left to right, in the top row the configurations with $\mathscr{L}_{\text{connectivity},SR}$ and $\mathscr{L}_{\text{connectivity},clDice}$, and in the bottom row the configurations with $\mathscr{L}_{\text{connectivity},cbDice}$ and $\mathscr{L}_{\text{connectivity},clCE}$. For each configuration, the curves depict the evolution of training loss (blue), validation loss (red), and validation Dice (green) over 1000 epochs.

faster but less accurate skeletonisation algorithm, as proposed for example in [20], would likely reduce the computational cost but may further degrade connectivity performance.

Beyond the per-epoch computational cost, the convergence behavior provides additional insight into the optimization dynamics of the connectivity-preserving losses. As shown in Figure 4.9, the $\mathscr{L}_{\text{connectivity},clDice}$ and $\mathscr{L}_{\text{connectivity},clCE}$ configurations reach a plateau in validation Dice after roughly 900 epochs, indicating full convergence within the allocated training budget. In contrast, both $\mathscr{L}_{\text{connectivity},SR}$ and, more pronounced, $\mathscr{L}_{\text{connectivity},cbDice}$ still exhibit a noticeable upward trend at epoch 1000, suggesting that they have not fully converged. This observation is particularly relevant for $\mathscr{L}_{\text{connectivity},cbDice}$, as it may still benefit from extended training, albeit at the cost of a markedly increased computational burden.

Taken together, the results show that all connectivity-preserving losses maintain comparable volumetric overlap performance relative to the generic baseline. Among them, $\mathscr{L}_{\text{connectivity},SR}$ provides the most consistent improvements in connectivity metrics, yields more complete distal vessel courses, and reduces the size of residual gaps, while incurring only a marginal

increase in training time. In contrast, the alternative connectivity-preserving terms offer weaker connectivity gains and are associated with a substantially higher computational cost. These findings identify the SR loss as the most effective and practical connectivity-preserving extension of the generic loss. Consequently, $\mathscr{L}_{\text{connectivity},SR}$ is selected as the representative connectivity-aware loss for the final experiment in Section 4.5, thereby addressing the central research question of this work.

4.4 Ablation Studies

This section investigates three design choices that may influence coronary artery tree segmentation performance. First, we evaluate whether adding a Difference of Gaussian (DoG)-based vesselness map as an auxiliary input channel improves overlap and connectivity metrics in Section 4.4.1. Second, we study how different recall targets (thin skeleton, tubed skeleton, full mask) affect the behavior of $\mathscr{L}_{\text{connectivity},SR}$ in Section 4.4.2. Third, we analyze postprocessing choices, including postprocessing-time cropping to avoid anisotropic patch geometry and lightweight connected-component filtering in Section 4.4.3.

4.4.1 Vesselness Filter

Several top-performing approaches in the ASOCA challenge employ vesselness filters as an additional input channel [27]. Vesselness filters are multiscale image operators that assign high responses to tubular, vessel-like structures while suppressing background. This suggests that explicitly enhancing tubular structures may facilitate coronary artery segmentation, particularly in low-contrast distal regions. We therefore investigated whether adding a vesselness map as an auxiliary input improves segmentation performance in our setting as well. Furthermore, we assess whether input-level vessel enhancement can partially compensate for the absence of a connectivity-preserving loss term.

To isolate the effect of vesselness-based input encoding, all models were trained using the same setup as in Section 4.3 and only employ a vesselness filter as an additional input channel to the CCTA intensities. The vesselness map was computed using a multiscale DoG-based filter. To incorporate anatomical prior knowledge, the analysis was restricted to tubular structures that are connected to the aorta. As a result, many spurious vessel-like responses that are not attached to the aortic lumen are discarded, which is expected to reduce the number of FP candidates.
To assess the impact of DoG-based vessel enhancement as an additional feature input on segmentation performance, we evaluated both $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ with and without the vesselness channel. In addition to the previously used overlap and connectivity metrics, we report the Positive Predictive Value (PPV) to explicitly quantify the burden of FP predictions. PPV corresponds to the precision term in the Dice formulation (see Equation 3.16). The results are summarized in Table 4.8.

Metric	CCTA		CCTA + Vesselness	
	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	78.76 [73.32, 81.42]	78.31 [74.86, 81.25]	78.95 [74.46, 81.41]	79.22 [74.51, 81.00]
PPV	81.13 [74.00,85.03]	76.52 [69.65,80.38]	80.72 [74.26, 84.14]	76.26[69.80, 79.94]
clDice	87.32 [83.92, 90.26]	86.95 [82.81, 89.34]	87.67 [83.30,90.65]	87.04 [83.13, 89.37]
εDice	93.58 [91.53, 95.71]	93.61 [91.18, 95.38]	93.66 [91.96, 95.71]	93.60 [91.74, 95.25]
clTPR	89.32 [83.88, 93.50]	93.01 [88.21, 96.84]	90.35 [84.34, 95.09]	93.82 [89.04, 97.69]
TPβ_0E	2 [1, 3]	1 [0, 3]	2 [1, 3]	1 [0, 3]

Table 4.8: Ablation of DoG-based vesselness input under the $\mathscr{L}_{\text{generic}}$ and the $\mathscr{L}_{\text{connectivity,SR}}$. Values denote median [Q1, Q3] over all cases.

Across all metrics, the inclusion of the vesselness map does not lead to consistent improvements. Dice, εDice, and clDice remain largely unchanged, indicating that the network already captures the relevant multiscale tubular appearance from the raw CCTA intensities alone. Interestingly, PPV systematically decreases for both loss configurations, showing that the vesselness channel tends to introduce additional FP predictions despite the aortic connectivity prior. A plausible explanation is that the vesselness map was added as an auxiliary input channel using the same normalization as the CCTA intensities. As a consequence, coronary arteries and other vessel-like structures that are connected to the aorta can obtain similar vesselness values. This reduces the network's ability to distinguish true coronary branches from neighbouring arterial or tubular structures, which is reflected in a drop in PPV.

While connectivity metrics increase slightly for both loss configurations when adding the vesselness channel, the $\mathscr{L}_{\text{generic}}$ with vesselness still does not reach the connectivity levels achieved by the $\mathscr{L}_{\text{connectivity},SR}$ without vesselness. This indicates that input-level vessel enhancement cannot substitute for an explicit connectivity-preserving loss term.

Overall, these findings show that DoG-based vessel enhancement does not provide a measurable benefit in our setting and cannot reproduce the connectivity gains achieved with $\mathscr{L}_{\text{connectivity},SR}$.

4.4.2 Tubed Skeleton

As described in Section 3.3.2, we follow the original mask transformation proposed in [21] to obtain the skeleton used for the SR loss calculation. Here, the one-voxel-wide skeleton is dilated with a diamond-shaped structuring element to form a tubular skeleton, which results in a three-voxel-wide representation of the reference mask. Given the voxel spacing of approximately 0.4 mm after preprocessing, a three-voxel-wide skeleton corresponds to an effective thickness of about 1.2 mm in our dataset.

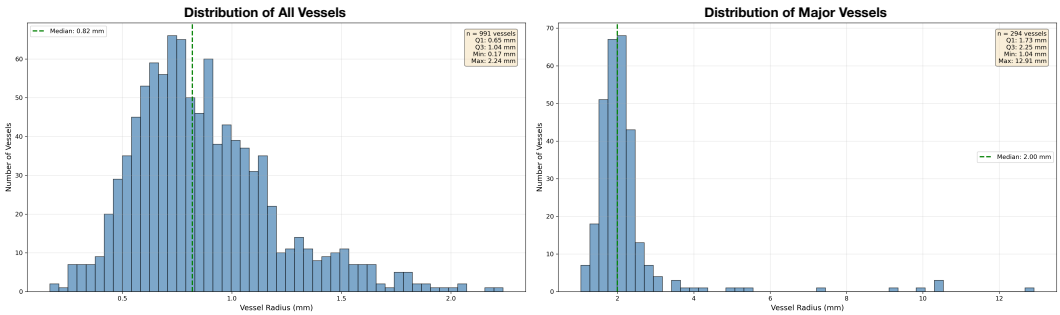

Figure 4.10: Distribution of vessel radii in the reference segmentations. The left panel shows the per-vessel median radius for all coronary vessels. The right panel shows the corresponding distribution restricted to the major vessels (RCA, LAD, and LCx),summarized by the per-vessel 95th percentile radius to reduce the influence of long, thin distal segments.

Metric	Thin skeleton	Tubed skeleton	Full mask
Dice	78.28 [73.94, 81.11]	78.31 [74.86, 81.25]	76.68 [72.89, 79.91]
clDice	85.73 [81.41, 88.59]	86.95 [82.81, 89.34]	87.63 [83.02, 90.23]
εDice	94.23 [92.38, 96.56]	93.61 [91.18, 95.38]	93.47 [90.89, 95.17]
clTPR	93.66 [89.42, 96.79]	93.01 [88.21, 96.84]	93.16 [88.45, 96.28]
TPβ_0E	2 [1, 3]	1 [0, 3]	1 [0, 2]

Table 4.9: Ablation of different recall targets for the SR loss. The thin skeleton is the one-voxel-wide skeleton, the tubed skeleton is the dilated three-voxel-wide variant used in the original formulation, and the full mask variant uses the full binary reference mask. Values denote median [Q1, Q3] over all cases.

Since the vast majority of coronary arteries have radii below this value, as shown in the left panel of Figure 4.10, the tubed skeleton occupies almost the entire extent of the reference mask for most vessels. Consequently, the transformed skeleton becomes highly similar to the original reference mask, with noticeable differences remaining only for the major vessels whose radii exceed 1.2 mm, as illustrated in the right panel of Figure 4.10.

Given this, we investigate how different recall targets affect the behavior of the SR loss. To this end, we compare three variants of $\mathscr{L}_{\text{connectivity},SR}$ that differ only in the definition of the recall mask used for the SR loss term: a thin skeleton without tubing, the original tubed skeleton, and the full binary reference mask. The results are summarized in Table 4.9.

Across the Dice metric, the thin skeleton and the tubed skeleton yield almost identical performance, whereas the full-mask variant shows a noticeable decrease. For the vessel accuracy metrics, all recall targets differ by less than 1 pp. Interestingly, the thin skeleton achieves the highest clTPR but shows an increased TPβ_0E compared to the other variants. The full-mask variant has the same median TPβ_0E as the tubed skeleton, but fewer cases with pronounced component fragmentation, as indicated by its lower upper quartile.

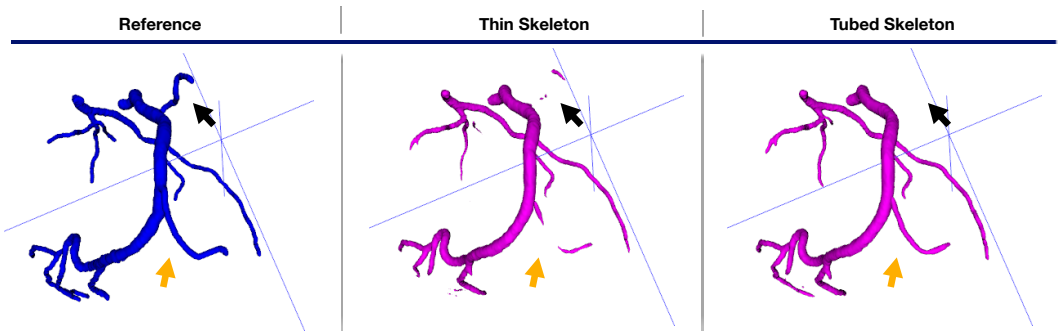

Figure 4.11: Qualitative comparison of recall targets for the SR loss. All results are visualized as 3D-rendered volumes. Column 1 shows the reference segmentation (blue). Column 2 shows the TP mask (magenta) obtained with the thin skeleton as recall target. Column 3 shows the TP mask (magenta) obtained with the tubed skeleton. Arrows highlight exemplary locations where the recall target affects local TP coverage and continuity.

A plausible explanation for these trends is the spatial support of the SR recall term, i.e., how widely it supervises the foreground and thereby controls the extent of its gradient signal. The thin, one-voxel-wide skeleton provides a highly sparse supervision signal that primarily enforces skeleton coverage, which is consistent with its slightly higher clTPR. However, the limited support around the thin skeleton suggests that voxels in immediate proximity to the skeleton do not consistently receive a recall-driven gradient. Consequently, the supervision provided by the SR term can be too spatially limited to reliably enforce continuity in the surrounding foreground, such that TP regions fragment into multiple components and TPβ_0E increases. In contrast, tubing the skeleton increases the spatial support of the recall term by extending the supervised band around the vessel axis. This makes the SR loss less sensitive to minor discretization effects and one-voxel misalignments, as voxels adjacent to the skeleton now also contribute to the recall signal. As a result, the recall-driven gradients provide a more coherent stabilizing signal for the surrounding foreground, which helps preserve local continuity and reduces component fragmentation, as reflected by the lower median TPβ_0E. This difference in spatial support is illustrated in Fig. 4.11, where the black arrow highlights a vessel course for which some voxels are recovered with the thin-skeleton target. However, this TP signal is fragmented into multiple components. In contrast, the corresponding vessel course is completely absent in the TP mask for the tubed-skeleton variant.

Notably, since the tubed skeleton already covers a large fraction of the reference mask in our data, the full-mask target does not primarily add geometric support. Instead, it changes the selectivity of the recall term by providing dense supervision across the entire reference mask. This can further suppress extreme TP fragmentation (lower upper quartile of TPβ_0E) but also biases the optimization towards higher foreground recall at the expense of precision, which is consistent with the reduced Dice. This shift in precision–recall behavior is reflected in Fig. 4.12, where the full-mask variant achieves the highest True Positive Rate (TPR) but shows a noticeable drop in PPV compared to the skeleton-based targets.

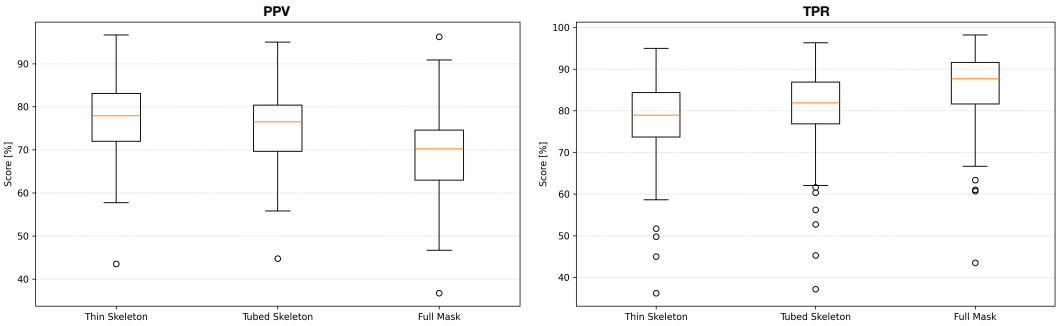

Figure 4.12: Precision–recall characteristics for different recall targets in the SR loss. Boxplots show PPV (left) and TPR (right) for the thin skeleton, tubed skeleton, and full-mask variants.

Overall, the tubed skeleton provides the most suitable recall target for our setting, as it offers a robust trade-off between overlap accuracy and connectivity preservation. At the same time, this ablation illustrates that even for segmentation problems where connectivity is the primary objective, volumetric overlap remains an important constraint. More generally, increasing the support of the recall target shifts the optimization towards higher foreground coverage and, in turn, higher centerline completeness, as the stronger gradient provides more supervision to the foreground. However, this also reduces the penalty for false positives, leading to lower precision and, consequently, a reduced Dice score. Therefore, the recall target should be chosen in a task-dependent manner: it should provide sufficient spatial support to reliably cover the relevant tubular structures—which can improve apparent continuity by reducing missed segments—while remaining selective enough to avoid predicting foreground beyond the true vessel extent.

4.4.3 Postprocessing

In Section 4.3, we used our heart-mask-based cropping strategy (described in Section 3.4.1.1) to accelerate training, which also affects the nnU-Net dataset fingerprint computed during its planning stage and thus the resulting patch size. In our case, planning on the cropped volumes led to a patch size of $160 \times 128 \times 112$. Since 3D CNN segmentation performance may deteriorate on anisotropically sampled data [84], we additionally performed the nnU-Net fingerprint extraction on the full volumes, which yields a patch size of $160 \times 160 \times 96$. Empirically, inference on full volumes produced FP predictions near the borders of the original field of view, i.e., outside the anatomical ROI. To mitigate these border artifacts, we applied the heart-mask-based cropping as a postprocessing step. The protocol is described in Section 3.4.4.

We first analyzed the effect of the cropping strategy on segmentation performance. Specifically, we compared the configuration trained with the anisotropic patch geometry induced by preprocessing-time heart-mask cropping with a configuration trained with an in-plane isotropic patch geometry, where heart-mask cropping was applied during postprocessing.

Metric	Pre-crop		Post-crop	
	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	78.76 [73.32, 81.42]	78.31 [74.86, 81.25]	79.11 [74.19, 81.46]	77.99 [74.40, 81.38]
clDice	87.32 [83.92, 90.26]	86.95 [82.81, 89.34]	87.58 [84.46, 90.78]	86.71 [82.91, 89.14]
εDice	93.58 [91.53, 95.71]	93.61 [91.18, 95.38]	94.17 [91.77, 95.92]	93.58 [91.43, 95.24]
clTPR	89.32 [83.88, 93.50]	93.01 [88.21, 96.84]	89.89 [84.49, 95.41]	93.86 [90.54, 97.60]
TPβ_0E	2 [1, 3]	1 [0, 3]	2 [1, 3]	1 [0, 2]

Table 4.10: Ablation of the heart-mask cropping strategy under $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,SR}}$. In the Pre-crop setting, cropping is applied during preprocessing, resulting in anisotropic patch geometry during training. In the Post-crop setting, cropping is applied during postprocessing, with training performed using an in-plane isotropic patch geometry. Values denote median [Q1, Q3] over all cases.

All other training settings were kept identical. For both configurations, we report results for $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ in Table 4.10.

Across the volumetric overlap metrics, applying cropping during postprocessing slightly improved performance under $\mathscr{L}_{\text{generic}}$, whereas under $\mathscr{L}_{\text{connectivity},SR}$ the overlap scores decreased marginally but remained within a comparable range. In contrast, clTPR consistently improved for both loss configurations in the post-crop setting. The TPβ_0E remained unchanged for $\mathscr{L}_{\text{generic}}$, while for $\mathscr{L}_{\text{connectivity},SR}$ the upper quartile decreased, indicating fewer cases with pronounced fragmentation.

Taken together, these results suggest that postprocessing-time cropping improves connectivity related metrics without compromising volumetric overlap. Consequently, we adopt the post-crop configuration for all subsequent connected component filtering experiments.
Connected component analysis is a common postprocessing step for segmentations of thin tubular structures, as it can suppress small spurious clusters and improve the apparent connectivity of the predicted tree. In contrast to topology-reconstruction approaches that explicitly reconnect or reconstruct vessel trees (e.g. [19]), which would add substantial computational overhead, we focus on lightweight postprocessing variants and evaluate whether they already improve segmentation quality.
We evaluated two connected-component-based postprocessing variants. In one variant, all connected components smaller than 100 voxels were removed and treated as noise artifacts. In the other variant, only the two largest connected components were retained, corresponding to the left and right coronary trees. The exact filtering settings are described in Section 3.4.4. To analyze the effect of postprocessing, we apply both connected-component filtering variants to the predictions obtained with the post-crop configuration. Empirically, many of the small connected components correspond to floating FP clusters. Therefore, removing components smaller than 100 voxels is expected to have only a minor impact on the TP-mask-based centerline completeness metrics reported in this work. Moreover, since these components contain only a small number of voxels, the effect on volumetric overlap metrics is expected to be negligible. To explicitly quantify the effect of filtering on the predicted mask, we

Metric	None	Size	Largest
Dice	79.11 [74.19, 81.46]	78.99 [74.19, 81.43]	78.75 [73.44, 81.71]
clDice	87.58 [84.46, 90.78]	87.66 [84.25, 90.47]	87.48 [83.18, 90.25]
εDice	94.17 [91.77, 95.92]	94.01 [91.74, 95.65]	93.69 [89.49, 95.65]
clTPR	89.89 [84.49, 95.41]	89.81 [83.91, 95.14]	84.01 [78.43, 92.35]
TPβ_0E	2 [1, 3]	2 [1, 3]	0 [0, 0]
β_0E	6 [4, 9]	4 [2, 6]	0 [0, 0]

Table 4.11: Ablation of connected-component-based postprocessing under $\mathscr{L}_{\text{generic}}$. In the None setting, no connected component filtering is applied. In the Size setting, components smaller than 100 voxels are removed. In the Largest setting, only the largest connected components are retained. Values denote median [Q1, Q3] over all cases.

Metric	None	Size	Largest
Dice	77.99 [74.40, 81.38]	77.98 [73.98, 81.27]	79.28 [75.24, 82.11]
clDice	86.71 [82.91, 89.14]	86.40 [82.73, 89.28]	87.76 [84.47, 90.79]
εDice	93.58 [91.43, 95.24]	93.35 [91.07, 95.18]	94.21 [92.09, 96.26]
clTPR	93.86 [90.54, 97.60]	93.62 [90.35, 97.43]	92.33 [84.64, 95.18]
TPβ_0E	1 [0, 2]	1 [0, 2]	0 [0, 0]
β_0E	6 [4, 8]	3 [2, 5]	0 [0, 0]

Table 4.12: Ablation of connected-component-based postprocessing under $\mathscr{L}_{\text{connectivity,SR}}$. In the None setting, no connected component filtering is applied. In the Size setting, components smaller than 100 voxels are removed. In the Largest setting, only the largest connected components are retained. Values denote median [Q1, Q3] over all cases.

additionally report the global β_0E. Table 4.11 reports the results under $\mathscr{L}_{\text{generic}}$, whereas Table 4.12 reports the results under $\mathscr{L}_{\text{connectivity},SR}$.

Beyond the expected negligible effect of removing small connected components, volumetric overlap metrics remain largely unchanged when retaining only the largest components for both loss configurations. This suggests that the proximal parts of the coronary trees are captured reliably in the predictions under both losses. However, under $\mathscr{L}_{\text{generic}}$, overlap metrics decrease slightly when retaining only the largest components, suggesting that distal vessel segments and small branching vessels are frequently not connected to the retained main components and are therefore removed by the filtering. This behavior is consistent with the pronounced drop in clTPR, whose median decreases from 89.89 to 84.01, indicating reduced centerline coverage for these thin, peripheral branches and distal vessel segments. In contrast, under $\mathscr{L}_{\text{connectivity},SR}$, overlap metrics increase when retaining only the largest components, while clTPR decreases only marginally from 93.86 to 92.33. This suggests that the removed components mainly represent floating FPs clusters rather than disconnected

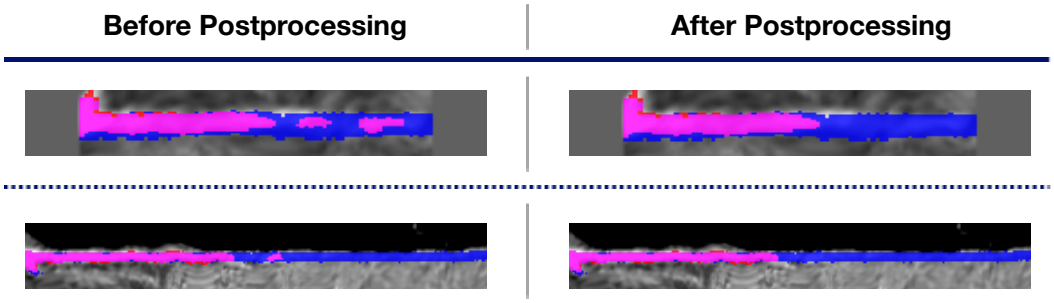

Figure 4.13: Qualitative examples illustrating the effect of connected-component-based postprocessing. Two cases are shown before (left) and after (right) connected-component filtering. Reference segmentations are shown in blue, and voxels where reference and prediction overlap are shown in magenta.

distal and peripheral vessel segments. This interpretation is supported by the reduction of the β_0E, whose median decreases from 6 to 3. For both loss configurations, TPβ_0E becomes 0 after retaining only the largest components, which is expected since the number of retained components is chosen to match the number of reference connected components.

Overall, filtering out small connected components can be considered a viable lightweight postprocessing step, as it has little, if any, adverse effect on volumetric overlap and centerline completeness for both loss configurations while reducing the global connected-component error. Nevertheless, the small performance drop observed for some metrics indicates that size-based filtering may also remove correctly predicted vessel parts. In particular, true vessel segments may appear as small components and can then be discarded, as illustrated by two examples in Figure 4.13. In these cases, the discarded component is separated from the main tree by only a small gap. A simple reconnection heuristic may therefore be preferable, e.g. by linking components that are spatially close along the expected vessel course [85], thereby preserving more of the coronary artery tree.

The same caveat applies when retaining only the largest components. This strategy requires prior information about the number of connected components for the given case. In most patients, retaining the two largest components is sufficient, as they typically correspond to the left and right coronary trees. However, anatomical variants such as multiple ostia, as observed in our cohort, can yield more than two major trees. An overly restrictive choice may therefore remove an entire correctly predicted vessel tree. If such case-specific information is available and this postprocessing is to be used, it should be applied only when the predicted coronary tree is already largely connected. Otherwise, in fragmented predictions, retaining only the largest components may discard correctly predicted vessel segments. This behavior is reflected in the strong decrease in clTPR observed under $\mathscr{L}_{\text{generic}}$, whereas under $\mathscr{L}_{\text{connectivity},SR}$ the decrease is only marginal. Since the main goal of this work is to compare loss configurations rather than to optimize the overall segmentation performance, we do not apply connected-component filtering in the final experiment (Section 4.5), as it would introduce an additional factor that disproportionately benefits $\mathscr{L}_{\text{connectivity},SR}$.

4.5 Final Experiment

In the final experiment, we statistically compare the two loss configurations $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ to address the research hypothesis. The comparison is conducted via 5-fold cross-validation on 98 cases. We use the nnU-Net-based evaluation pipeline with the parameter settings and processing steps established in the preceding sections. Volumetric overlap is quantified using Dice for vessel mask accuracy as well as clDice and εDice for vessel accuracy. Connectivity is evaluated using clTPR and TPβ_0E for centerline completeness, complemented by GC to quantify residual gaps. Computational cost is measured by the total training time. In light of these evaluation criteria, the research hypothesis is defined as:

Research Hypothesis

Adding Skeleton Recall to the generic loss improves coronary artery tree connectivity compared to the generic loss alone, without substantially degrading volumetric accuracy or increasing computational cost.

To formally test this hypothesis, we formulate the following null hypotheses:

Null Hypotheses

- H_0^{vol}: $\mathscr{L}_{\text{connectivity},SR}$ substantially degrades volumetric overlap compared to $\mathscr{L}_{\text{generic}}$.
- H_0^{conn}: $\mathscr{L}_{\text{connectivity},SR}$ does not improve coronary artery tree connectivity compared to $\mathscr{L}_{\text{generic}}$.
- H_0^{comp}: $\mathscr{L}_{\text{connectivity},SR}$ substantially increases computational cost compared to $\mathscr{L}_{\text{generic}}$.

As shown in Section 4.2.1, the evaluation metrics exhibit predominantly non-normal distributions. We therefore employ non-parametric statistical tests. Specifically, we use one-sided Wilcoxon signed-rank tests at a significance level of $\alpha = 0.05$, as detailed in Section 3.6. To test H_0^{vol} and H_0^{comp}, we perform non-inferiority tests with margins of $\delta_{\text{vol}} = 1$ and $\delta_{\text{comp}} = 3$, corresponding to one percentage point in volumetric overlap and approximately 10% of the training time observed for $\mathscr{L}_{\text{generic}}$, respectively. We choose one percentage point as a practical equivalence margin, since differences of this magnitude are typically within the expected variability of overlap metrics and rarely reflect a meaningful qualitative change. We consider a 10% increase in training time a modest and acceptable overhead, as training is performed offline and does not affect clinical turnaround, provided that it yields measurable improvements in coronary tree connectivity. To test H_0^{conn}, we apply a superiority test to assess whether $\mathscr{L}_{\text{connectivity},SR}$ achieves significantly higher connectivity.

For volumetric overlap and connectivity, Table 4.13 reports the per-metric results together with the corresponding p-values obtained from paired statistical tests across 98 cases. For

Metric	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	p-value
Dice	79.11 [74.19, 81.46]	77.99 [74.40, 81.38]	9.15×10^{-4}
clDice	87.58 [84.46, 90.78]	86.71 [82.91, 89.14]	4.26×10^{-3}
εDice	94.17 [91.77, 95.92]	93.58 [91.43, 95.24]	2.53×10^{-4}
clTPR	89.89 [84.49, 95.41]	93.86 [90.54, 97.60]	2.61×10^{-17}
TP β_0 E	2 [1, 3]	1 [0, 2]	2.70×10^{-6}
GC	1 [0, 2]	1 [0, 1]	1.00×10^{-3}

Table 4.13: Comparison of volumetric overlap and connectivity metrics between $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,SR}}$. Values are reported as median [Q1, Q3] across 98 cases; p-values are obtained from paired one-sided Wilcoxon signed-rank tests.

Metric	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	p-value
Runtime	26.50 [25.47, 26.53]	29.28 [29.08, 29.85]	3.13×10^{-2}

Table 4.14: Comparison of training runtime in hours across the five paired cross-validation folds. Values are reported as median [Q1, Q3]; p-value is obtained from paired one-sided Wilcoxon signed-rank tests.

computational cost, Table 4.14 reports the results and the corresponding p-value obtained from paired tests across the five cross-validation folds.

Overall, the results show that adding the SR leads to clear improvements in coronary artery tree connectivity, accompanied by only minor reductions in volumetric overlap and a modest increase in training time. All observed effects were statistically significant under the applied testing framework, either in terms of superiority or non-inferiority.
Accordingly, we reject H_0^{vol}, H_0^{conn}, and H_0^{comp} and accept following alternative hypotheses:

Alternative Hypotheses

- H_A^{vol}: $\mathscr{L}_{\text{connectivity},SR}$ does not substantially degrade volumetric overlap compared to $\mathscr{L}_{\text{generic}}$.
- H_A^{conn}: $\mathscr{L}_{\text{connectivity},SR}$ improves coronary artery tree connectivity compared to $\mathscr{L}_{\text{generic}}$.
- H_A^{comp}: $\mathscr{L}_{\text{connectivity},SR}$ does not substantially increase computational cost compared to $\mathscr{L}_{\text{generic}}$.

To put the aggregate findings into a practical perspective, we complement the quantitative comparison with qualitative case studies. As a first step, we illustrate representative extreme cases by selecting the best and worst cases according to the Dice score across both loss configurations. Table 4.15 reports the corresponding quantitative metrics, whereas Figures 4.14 and 4.15 provide qualitative comparisons based on rendered 3D volumes and sMPRs, respec-

Metric	Worst case		Best case	
	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$	$\mathscr{L}_{\text{generic}}$	$\mathscr{L}_{\text{connectivity,SR}}$
Dice	51.75	53.44	87.12	87.24
εDice	95.80	95.75	98.16	97.80
TPβ_0E	3	0	4	0

Table 4.15: Quantitative metrics for the global best and worst cases selected according to the Dice score across $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,SR}}$.

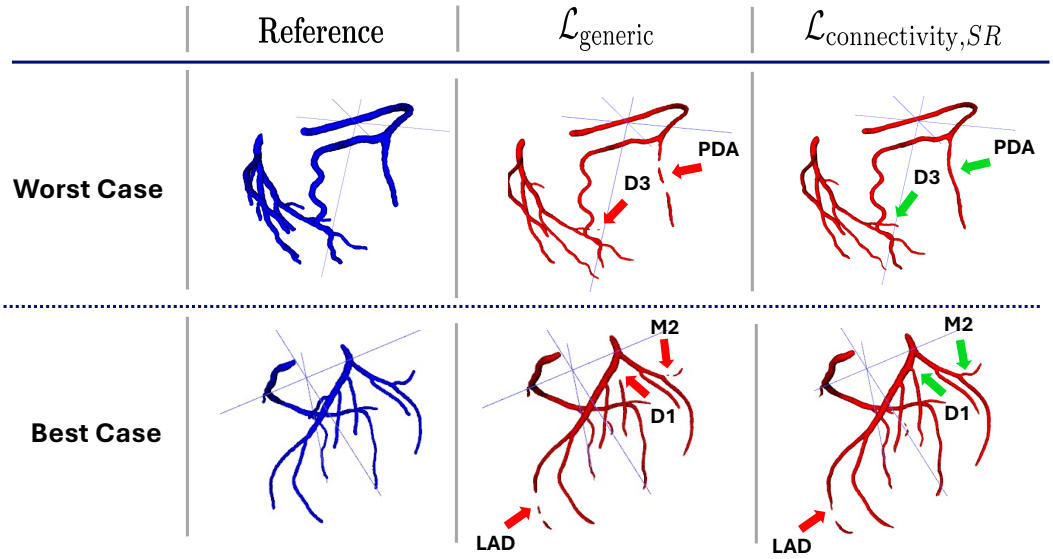

Figure 4.14: Qualitative comparison of the global worst (top row) and best (bottom row) cases selected according to the Dice score across $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$. The reference is shown in blue and the predictions in red. All results are visualized as 3D-rendered volumes. Column 1 shows the reference, Column 2 the prediction obtained with $\mathscr{L}_{\text{generic}}$, and Column 3 the prediction obtained with $\mathscr{L}_{\text{connectivity},SR}$. Red arrows highlight disconnected vessel segments, whereas green arrows indicate segments that remain connected.

tively. The sMPRs focus on vessel segments that exhibit visible connectivity breaks in the TP mask, allowing a direct comparison between the corresponding segmentations obtained with $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$.

The quantitative metrics of the selected cases are consistent with the global statistical findings. While differences in volumetric overlap are small in magnitude, connectivity-related metrics show pronounced improvements under $\mathscr{L}_{\text{connectivity},SR}$. For the worst case, Dice is substantially lower than εDice for both loss configurations, indicating that most discrepancies arise from boundary inaccuracies and locally reduced vessel thickness rather than from failures in capturing the underlying vessel topology. This behavior is also visible in the rendered volumes, where the reference annotation appears noticeably thicker than both

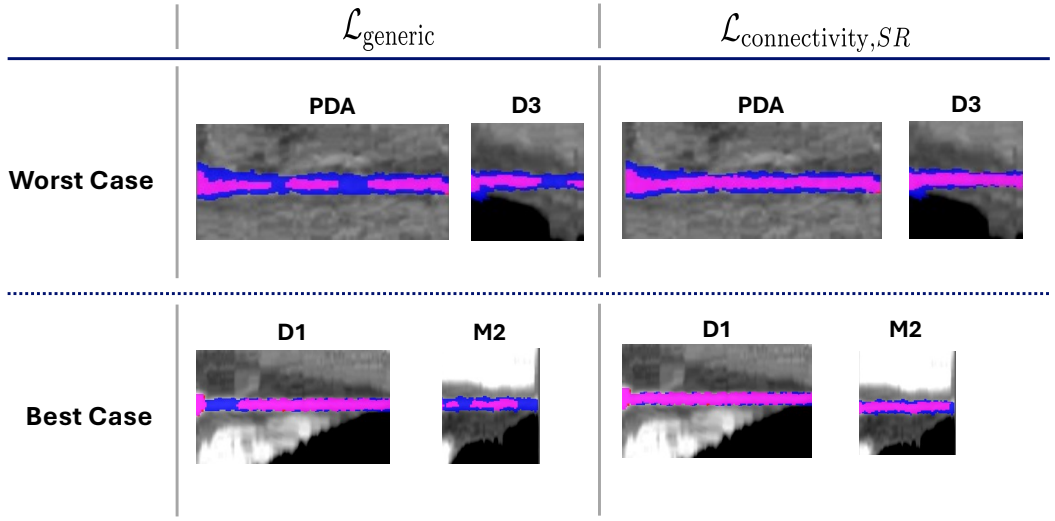

Figure 4.15: sMPRs of the vessel segments highlighted by arrows in Figure 4.14, shown for both $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$. The reference segmentation is shown in blue, and voxels where reference and prediction overlap are shown in magenta.

predictions. The sMPRs further reveal incomplete boundary coverage along representative vessel segments.

While these boundary deviations largely explain the low Dice in this case, they do not account for the main qualitative difference between both loss configurations, which lies in the connectivity of the segmented tree. Specifically, under $\mathscr{L}_{\text{generic}}$ several distal vessel segments are fragmented, resulting in multiple disconnected TP components. These breaks are highlighted by the red arrows in the rendered volumes and are also evident in the corresponding sMPRs, where the vessel course is interrupted. In contrast, $\mathscr{L}_{\text{connectivity},SR}$ preserves these segments as a continuous vessel course, yielding a TPβ_0E of 0.

For the best case, Dice and εDice are closely aligned for both loss configurations. Qualitatively, this is reflected in the rendered volumes, where the predictions follow a similar vessel course as the reference. Compared to the worst case, the reference annotation exhibits a less pronounced vessel thickness. Consistently, the sMPRs show a more complete coverage of the vessel boundaries.

Nevertheless, connectivity-related differences also persist in this case. While $\mathscr{L}_{\text{generic}}$ still exhibits small discontinuities in distal branches, $\mathscr{L}_{\text{connectivity},SR}$ again yields a fully connected true-positive mask, reflected by a TPβ_0E of 0.

Notably, along the LAD a distal connectivity break can still be observed under $\mathscr{L}_{\text{connectivity},SR}$. This break occurs beyond the annotated reference extent and is therefore classified as distal FP, and thus does not affect TPβ_0E. Interestingly, we observed this pattern in many cases: the prediction extends beyond the end of the reference annotation and subsequently fragments shortly after the reference terminates. To systematically identify and quantify this and related error patterns, we analyze the proposed subclasses of FP and FN voxels introduced in Section 3.5.6. An example of this distal FP fragmentation pattern is shown in Figure 4.16.

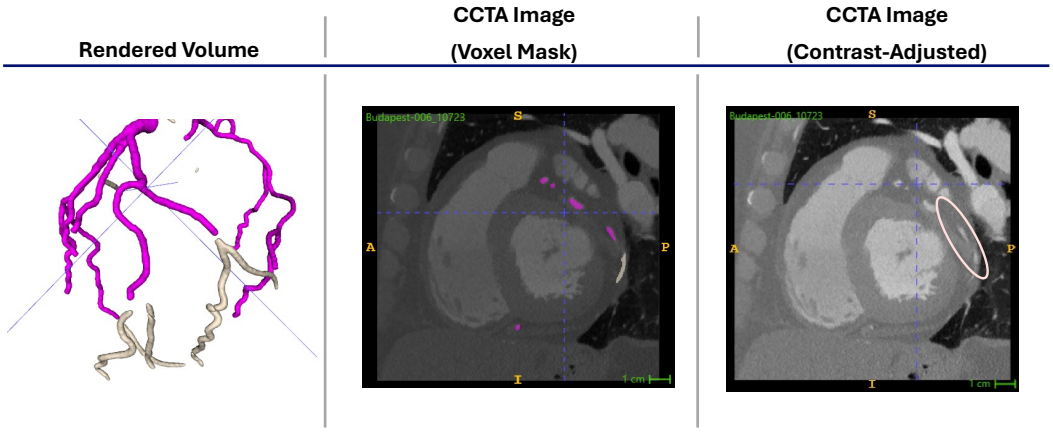

Figure 4.16: Example of a case with distal FP fragmentation beyond the annotated reference extent under $\mathscr{L}_{\text{connectivity},SR}$. The prediction is decomposed into TP voxels (magenta) and $\mathbf{FP}_{\text{dist}}$ voxels (beige). Column 1 shows the 3D-rendered volume of the decomposed prediction. Column 2 shows a corresponding CCTA slice with voxel overlay. Column 3 shows a contrast-adjusted CCTA slice; the beige ellipse highlights an anatomical region contributing to the $\mathbf{FP}_{\text{dist}}$ extension.

The prediction contains a noticeable proportion of $\mathbf{FP}_{\text{dist}}$ voxels extending beyond the annotated reference extent. The corresponding CCTA views suggest the presence of vessel-like structures with similar intensities beyond the annotation terminus, indicating that the reference may be truncated in this region. However, the distal continuation is not predicted as a single coherent course, such that a small gap within the extension leads to distal connectivity breaks.

Notably, the same distal fragmentation pattern is also observed under $\mathscr{L}_{\text{generic}}$, indicating that this behavior is not specific to the added SR loss. Instead, it appears to be a more general property of the network and the training data. In particular, when vessel-like structures are visible beyond the annotated reference extent, the network tends to extrapolate the vessel course based on local appearance cues. In the absence of a supervisory signal beyond the annotation boundary, such extrapolations can become unstable and fragment shortly after the reference's distal end.

Another typical FP pattern observed in the predictions, independent of the loss function, corresponds to the $\mathbf{FP}_{\text{float}}$ subclass. In many cases, these isolated components coincide not only with noise-driven artifacts but also with vessel-like structures visible in the CCTA volume, such as veins or other tubular anatomical structures. Figure 4.17 illustrates an example of this pattern.

In this example, the largest $\mathbf{FP}_{\text{float}}$ component has a tubular shape that resembles a typical coronary vessel. Similar to $\mathbf{FP}_{\text{dist}}$ extensions, the corresponding CCTA image exhibits vessel-like structures and contrast-enhancement patterns at this location, suggesting the presence of a vascular anatomical structure. In this particular case, the structure is consistent with a venous vessel rather than a coronary artery. This observation indicates that the network relies strongly on local appearance cues, which can lead to the segmentation of vessel-like tubular structures that do not belong to the coronary artery tree.

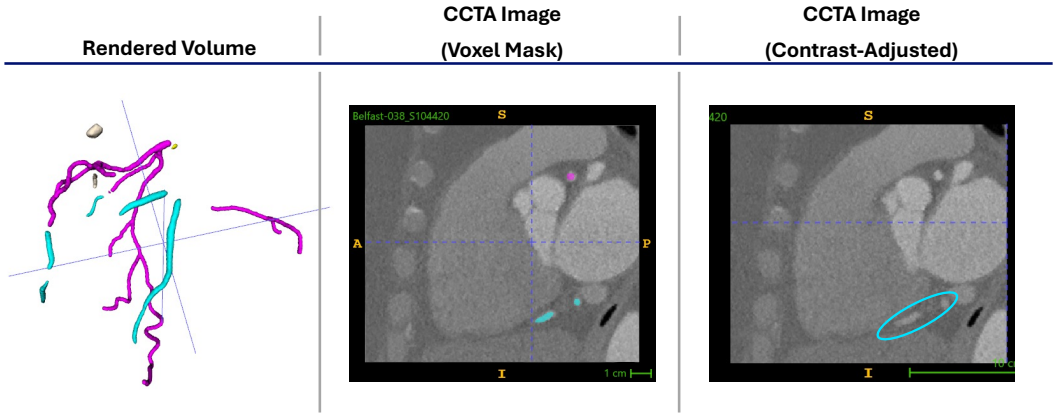

Figure 4.17: Example of a case with floating FP components. The prediction is decomposed into TP voxels (magenta) and $\mathbf{FP}_{\text{float}}$ voxels (cyan). Column 1 shows the 3D-rendered volume of the decomposed prediction. Column 2 shows a corresponding CCTA slice with mask overlay. Column 3 shows a contrast-adjusted CCTA slice; the cyan ellipse highlights an anatomical region associated with a $\mathbf{FP}_{\text{float}}$ component.

A plausible contributing factor is contrast timing. Although CCTA is optimized for arterial coronary opacification, veins can exhibit comparable attenuation depending on bolus timing, cardiac output, and patient-specific hemodynamics. As contrast passes through the cardiac chambers and subsequently opacifies the venous circulation, parts of the venous system may appear strongly enhanced at the time of acquisition, providing appearance cues that can trigger $\mathbf{FP}_{\text{float}}$ predictions. Moreover, limited training data may prevent the network from learning a strong spatial prior for the typical location of the coronary arteries.

The remaining subclass, $\mathbf{FP}_{\text{prox}}$, occurs only in two cases with a noticeable magnitude. Visual inspection suggests that these instances can be attributed to missing proximal annotations near the coronary ostium, which were subsequently corrected during data curation. Since the observed $\mathbf{FP}_{\text{prox}}$ pattern is thus not representative of the finalized dataset, we do not further analyze this subclass in the following.

Although $\mathscr{L}_{\text{connectivity},SR}$ improves connectivity significantly, a subset of cases still exhibits residual discontinuities in the predicted coronary artery tree. We therefore analyze subclasses of FN voxels to characterize systematic missing voxels relative to the reference segmentation. While our FN confusion set subdivides into $\mathbf{FN}_{\text{thin}}$ for locally reduced vessel thickness and $\mathbf{FN}_{\text{short}}$ for prematurely truncated predictions, the subclass $\mathbf{FN}_{\text{disc}}$ directly affects tree connectivity and is therefore particularly relevant in the context of this work.
We therefore focus on voxels classified as $\mathbf{FN}_{\text{disc}}$ that persist even under $\mathscr{L}_{\text{connectivity},SR}$ and thus represent failure modes that are not resolved by connectivity-aware supervision. Figure 4.18 illustrates a representative case in which an $\mathbf{FN}_{\text{disc}}$ segment corresponds to a missing vessel portion that would connect two otherwise disconnected TP components, thereby manifesting as a connectivity break in the network prediction. In the contrast-adjusted CCTA image, the proximal and distal parts of the vessel show clear contrast enhancement,

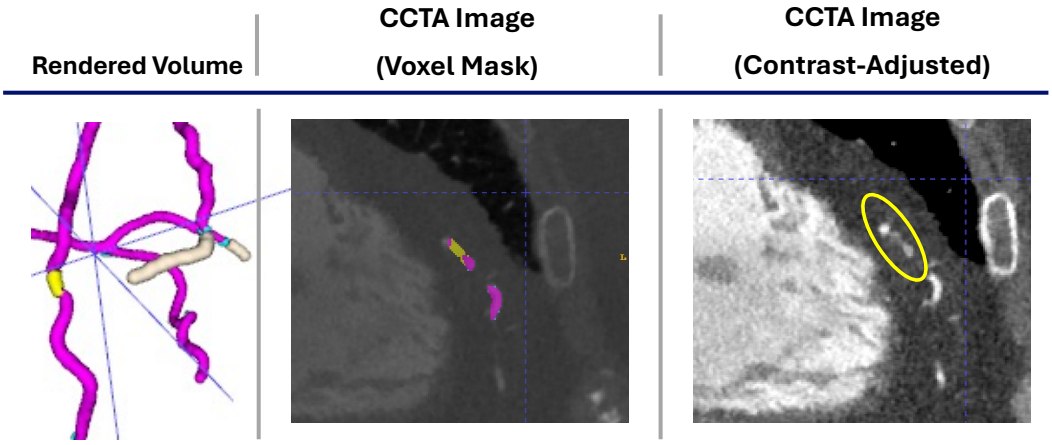

Figure 4.18: Example of a case with decomposed prediction for FN subclasses under $\mathscr{L}_{\text{connectivity},SR}$. The prediction is decomposed into TP voxels (magenta), $\mathbf{FN}_{\text{disc}}$ voxels (yellow), $\mathbf{FN}_{\text{thin}}$ voxels (cyan) and $\mathbf{FN}_{\text{short}}$ voxels (beige). Column 1 shows the 3D-rendered volume of the decomposed prediction. Column 2 shows a CCTA slice with voxel overlay focusing on the $\mathbf{FN}_{\text{disc}}$ region. Column 3 shows the corresponding contrast-adjusted CCTA slice; the yellow ellipse highlights the anatomical region associated with the $\mathbf{FN}_{\text{disc}}$ component.

	TP	$\mathbf{FN}_{\text{disc}}$	$\mathbf{FN}_{\text{thin}}$	$\mathbf{FN}_{\text{short}}$
HU	354.33	172.13	181.02	139.87

Table 4.16: Median HU values over all TP voxels and FN voxel subclasses in the cases under $\mathscr{L}_{\text{connectivity,SR}}$.

whereas the intermediate segment exhibits markedly reduced attenuation, coinciding with the region missed by the model. This observation is consistent with Table 4.16, where $\mathbf{FN}_{\text{disc}}$ voxels exhibit substantially lower median HU values than TP voxels.

Similar trends are observed for $\mathbf{FN}_{\text{thin}}$ and $\mathbf{FN}_{\text{short}}$, suggesting that FN predictions preferentially occur in vessel portions with reduced HU values relative to typically well-enhanced adjacent coronary segments. Notably, this pattern is observed irrespective of the loss configuration. In particular for $\mathbf{FN}_{\text{disc}}$, it indicates that locally reduced vessel contrast constitutes a key appearance-related bottleneck that persists even under connectivity-aware supervision. To further support this interpretation, we compare the predicted foreground softmax probability on $\mathbf{FN}_{\text{disc}}$ voxels and, for reference, on TP voxels under both $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ in Table 4.17.
Both losses assign high foreground probability to **TP** voxels, whereas $\mathbf{FN}_{\text{disc}}$ voxels consistently receive low foreground probability. This supports our interpretation that appearance-related limitations dominate these errors and are not substantially mitigated by the choice of loss configuration.

	$\mathscr{L}_{\text{generic}}$		$\mathscr{L}_{\text{connectivity,SR}}$	
	TP	$\mathbf{FN}_{\text{disc}}$	TP	$\mathbf{FN}_{\text{disc}}$
FG Probability	92	18	93	20

Table 4.17: Foreground (FG) softmax probabilities for TP and $\mathbf{FN}_{\text{disc}}$ under $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity,SR}}$. Values denote the median over all voxels in the respective set.

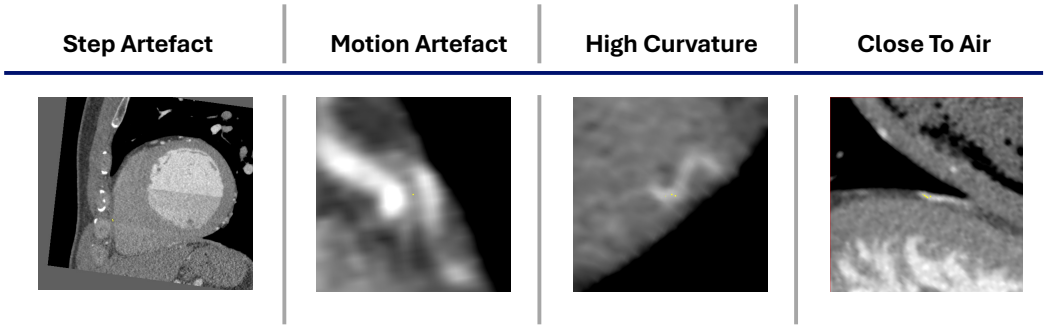

Figure 4.19: Representative examples of image-related root causes of connectivity disruptions.

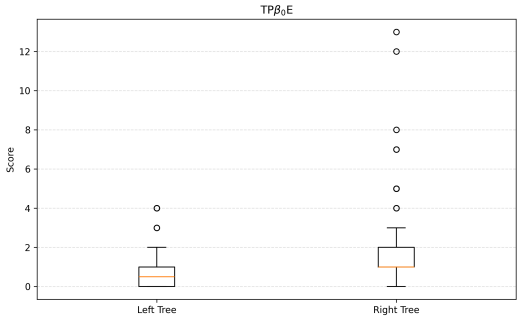

Figure 4.20: Boxplot of TPβ_0E for the left and right coronary artery tree under $\mathscr{L}_{\text{connectivity},SR}$.

Beyond low-contrast vessel portions, additional image-related factors were identified as recurrent root causes of connectivity disruptions in the predicted coronary artery tree that can still occur under $\mathscr{L}_{\text{connectivity},SR}$. Figure 4.19 illustrates representative examples, including step artifacts, motion-induced blurring, high vessel curvature, and vessel segments located close to air-filled structures.

In addition to the error pattern analysis, we observed systematic differences between the left and right coronary artery trees. As illustrated in Figure 4.20, the left coronary tree exhibits a lower median TPβ_0E and a comparatively compact interquartile range, whereas the right coronary tree shows a broader distribution with several pronounced outliers. This indicates that severe connectivity disruptions occur more frequently in the right coronary tree.

This indicates that severe connectivity disruptions occur more frequently in the right coronary tree. A plausible contributing factor is the use of a late-diastolic ECG-gated reconstruction window in our cohort. Late diastole often provides high image quality for the LAD and LCx, but can be suboptimal for the RCA and may increase sensitivity to motion- and reconstruction-related artifacts [86]. Furthermore, we observed that the reference annotations were more complete for the left coronary tree than for the right tree, which can affect both model supervision and the evaluation of connectivity-related errors and may therefore contribute to the observed differences.

Taken together, the qualitative case studies and the error decomposition suggest that the main failure modes are largely driven by appearance- and acquisition-related limitations in coronary artery segmentation. Nevertheless, the statistical analysis demonstrates that adding SR to the generic loss yields statistically significant and practically relevant improvements in coronary artery tree connectivity, while maintaining comparable volumetric overlap and incurring only a modest increase in training time. This makes SR a robust and lightweight approach for improving connectivity and topology preservation without relying on complex pipelines or extensive postprocessing.

Chapter 5

Conclusion

This chapter summarizes the experiments and discusses the main findings and their implications. It concludes by answering the research question posed in Section 1.3 and by outlining potential directions for future work.

The thesis follows a typical end-to-end machine learning workflow for medical image segmentation. It starts with data curation and dataset analysis, which involves validating the available cohort and refining it via consistency checks and quality control. A lightweight exploratory pipeline enables rapid iteration and low-overhead preliminary experiments, thereby validating key design choices. Building on these insights, an nnU-Net-based evaluation pipeline quantifies the benefits and limitations of connectivity-preserving losses compared to a generic baseline loss for coronary artery segmentation. Performance is assessed using complementary metrics covering vessel mask accuracy, vessel accuracy, centerline completeness, and runtime, thereby capturing volumetric overlap, connectivity-related effects, and computational cost. In addition, paired statistical tests determine whether observed differences between the generic baseline loss and the SR-based connectivity-preserving configuration are statistically significant.

The experimental results are summarized in the following. We first examine three sliding-window overlap configurations (0%, 25%, and 50%) in Section 4.1, focusing on their impact on connectivity and inference time. A 50% overlap yields the best connectivity but substantially increases inference time. However, since inference accounts for only 1.5% of the end-to-end runtime in this setup, this overhead is negligible in practice, and 50% overlap is adopted for all subsequent experiments.

The baseline experiments in Section 4.2 use the exploratory pipeline to provide an initial comparison between $\mathscr{L}_{\text{generic}}$ and $\mathscr{L}_{\text{connectivity},SR}$ and to establish a reference configuration for the subsequent nnU-Net-based evaluation pipeline.
The data distribution analysis in Section 4.2.1 shows that most evaluation metrics exhibit non-normal distributions, motivating the use of medians and interquartile ranges for reporting as well as non-parametric paired tests in the final statistical analysis.
The metric suitability analysis in Section 4.2.2 shows that HD_{95} is strongly influenced by peripheral FP structures and therefore poorly reflects vessel delineation quality in this setting.

Similarly, the commonly used β_0E is partly dominated by boundary-related FP components rather than reflecting true coronary tree connectivity. Consequently, Dice is adopted as the representative vessel mask accuracy metric, clDice and εDice are used as vessel accuracy metrics, and clTPR together with TPβ_0E are selected as the primary connectivity measures. Finally, the blueprint parameter study in Section 4.2.3 identifies the U-Net template with SGD and Nesterov momentum combined with the PolyLR learning-rate schedule as a stable configuration for the nnU-Net framework.

In Section 4.3, the nnU-Net-based evaluation pipeline systematically compares $\mathscr{L}_{\text{generic}}$ with multiple connectivity-preserving losses ($\mathscr{L}_{\text{connectivity},SR}$, $\mathscr{L}_{\text{connectivity},clDice}$, $\mathscr{L}_{\text{connectivity},cbDice}$, and $\mathscr{L}_{\text{connectivity},clCE}$).
Across all overlap-based metrics, including Dice, clDice, and εDice, performance remains largely invariant, indicating that connectivity-aware supervision does not compromise volumetric accuracy.
In contrast, the connectivity metrics exhibit pronounced differences between loss configurations. $\mathscr{L}_{\text{connectivity},SR}$ provides the most consistent improvement, reflected by higher clTPR and lower TPβ_0E, with favorable distribution shifts and fewer low-performing outliers. Qualitative analysis further confirms that the SR-based model yields more continuous coronary trees with more complete distal vessel courses, whereas other loss formulations frequently exhibit premature terminations or fragmentation.
In addition, computational analysis shows that $\mathscr{L}_{\text{connectivity},SR}$ incurs only a marginal increase in training time, while the other connectivity-preserving losses substantially increase computational cost.
Taken together, these results identify SR as the most effective and practically viable connectivity-preserving term in this setting.

The ablation studies in Section 4.4 investigate how selected design choices influence coronary artery tree segmentation.
Adding a multiscale DoG-based vesselness map as an auxiliary input channel in Section 4.4.1 does not yield consistent improvements in either overlap or connectivity metrics under $\mathscr{L}_{\text{generic}}$ or $\mathscr{L}_{\text{connectivity},SR}$, indicating that vesselness-based input augmentation cannot reproduce the connectivity gains provided by the SR term.
Section 4.4.2 compares a thin skeleton, a tubed skeleton, and the full reference mask as recall targets for the SR loss. The results show that larger spatial support improves centerline completeness but shifts the optimization toward higher foreground coverage, reducing precision. In this dataset, the tubed skeleton provides the most robust trade-off and is selected as the preferred recall target.
The postprocessing study in Section 4.4.3 shows that applying heart-mask cropping during postprocessing rather than during preprocessing improves connectivity-related metrics while maintaining comparable volumetric overlap, as it allows training with an in-plane isotropic patch geometry.
Beyond cropping, Section 4.4.3 also evaluates lightweight connected-component filtering as an additional postprocessing option. This filtering reduces the global connected-component

error but can remove correctly predicted distal segments, particularly for fragmented predictions under $\mathscr{L}_{\text{generic}}$, leading to reduced centerline completeness. Since it disproportionately benefits $\mathscr{L}_{\text{connectivity},SR}$, it is not applied in the final experiment to ensure a fair comparison.

Section 4.5 reports the final, hypothesis-driven experiment that formally addresses the research question. The analysis tests superiority for connectivity and non-inferiority for volumetric overlap and training time when comparing $\mathscr{L}_{\text{connectivity},SR}$ against $\mathscr{L}_{\text{generic}}$. The results show a clear and statistically significant improvement in connectivity, while the small reductions in volumetric overlap and the increase in training time remain within the predefined non-inferiority margins.
A recent manuscript [87] questions the general effectiveness of $\mathscr{L}_{\text{connectivity},SR}$ and reports that it does not consistently outperform the generic loss. However, the study does not include connectivity-aware evaluation metrics, which are essential for assessing thin tubular structure segmentation quality. As a result, conclusions drawn solely from volumetric overlap metrics do not fully capture the primary objective of adding the SR term, namely improved centerline completeness and fewer connectivity breaks, a trend that is consistent across our experiments and becomes statistically significant in the final analysis.

Despite these improvements, a subset of cases still exhibits residual errors and local discontinuities. One recurring pattern is distal extension beyond the annotated reference extent followed by fragmentation shortly after the annotation terminates, which is primarily driven by annotation truncation. Future work can investigate endpoint extension strategies or continuity-aware regularization terms that explicitly stabilize distal predictions beyond the annotated reference extent.
Another failure mode corresponds to vessel portions with reduced HU values relative to adjacent segments. For these voxels, the network assigns low foreground confidence, resulting in gaps along otherwise continuous vessel courses. Additional image-related factors, such as motion blur and high curvature, also contribute to disruptions. Improved robustness can be pursued by increasing training-data diversity with respect to anatomical variants and acquisition conditions and by emphasizing low-attenuation vessel voxels during training.
Besides connectivity-breaking errors, the network also segments non-coronary vessel-like structures, such as veins, as foreground. Potential mitigation strategies include additional supervision by labeling venous structures and training a multi-class model or adding a loss term that penalizes predictions in regions annotated as veins.

We conclude and directly answer the research question; the benefits and challenges of integrating connectivity-preserving losses depend strongly on their specific configuration. Across all investigated loss formulations, overlap-based performance remains comparable to the generic baseline, indicating that connectivity-aware supervision does not inherently degrade volumetric agreement. However, substantial differences arise in connectivity gains and computational cost. $\mathscr{L}_{\text{connectivity},clDice}$, $\mathscr{L}_{\text{connectivity},cbDice}$, and $\mathscr{L}_{\text{connectivity},clCE}$ rely on differentiable skeleton-based methods and incur pronounced increases in training time while providing limited or inconsistent improvements in connectivity-related metrics. In contrast, augmenting the generic loss with an SR term improves coronary artery tree connectivity in

a statistically significant and practically relevant manner while keeping the training-time overhead negligible. Its advantages stem from a simple and robust design: the loss requires only a precomputed reference skeleton, which can be generated during data loading using standard libraries. Furthermore, the approach is architecture-agnostic and straightforward to implement, providing a practical connectivity improvement without relying on complex pipelines or extensive postprocessing.

Beyond the scope of this work, several directions for future research emerge. A central aspect concerns annotation consistency across datasets. In this study, the two cohorts differ in annotation protocols and annotated extent, leading to systematic differences in distal vessel labeling and label thickness. These discrepancies bias both overlap- and skeleton-based evaluation and can induce apparent fragmentation at annotation boundaries. Standardizing labeling guidelines across cohorts, potentially supported by interactive annotation tools, would reduce such dataset-specific effects and enable a more reliable comparison of segmentation methods.
In addition, enlarging the cohort and rolling out the pipeline to larger multi-center datasets, such as ImageCAS [88], would allow for a more robust evaluation of generalization.

Another promising direction is the explicit handling of recurrent FP predictions. Model predictions can be used in an iterative, model-assisted workflow to identify hard negatives: repeatedly collecting high-confidence FP regions (e.g., veins or other vessel-like structures), manually annotating them as an additional negative mask, and fine-tuning the model with loss terms that penalize coronary predictions in these confounder regions. This hard-negative mining strategy could reduce spurious components. Complementarily, explicitly predicting a coronary-specific ROI, e.g., via a dedicated localization network or joint multi-task learning, could suppress FP predictions outside the coronary tree and simplify the subsequent segmentation task.

To address residual FNs and small connectivity breaks, future work can further investigate strategies that explicitly target gap closure and distal completeness. This includes postprocessing techniques such as centerline-guided reconnection or constrained morphological bridging, as well as training-time approaches that emphasize low-confidence vessel voxels in low-attenuation or artifact-affected segments.

References

[1] World Health Organization, "World health statistics 2025: monitoring health for the sdgs, sustainable development goals," World Health Organization, Geneva, Tech. Rep., 2025, global report. https://www.who.int/publications/i/item/9789240110496

[2] C. Vrints, F. Andreotti, K. C. Koskinas, X. Rossello, and et al., "2024 esc guidelines for the management of chronic coronary syndromes," *European Heart Journal*, vol. 45, pp. 3415–3537, 2024. doi:10.1093/eurheartj/ehae177

[3] M. J. Budoff, D. Li, E. A. Kazerooni, G. S. Thomas, J. H. Mieres, and L. J. Shaw, "Diagnostic accuracy of noninvasive 64-row computed tomographic coronary angiography (ccta) compared with myocardial perfusion imaging (mpi): the picture study, a prospective multicenter trial," *Academic Radiology*, vol. 24, pp. 22–29, 2017.

[4] A. J. Foy, S. S. Dhruva, B. Peterson, J. M. Mandrola, D. J. Morgan, and R. F. Redberg, "Coronary computed tomography angiography vs functional stress testing for patients with suspected coronary artery disease: a systematic review and meta-analysis," *JAMA internal medicine*, vol. 177, pp. 1623–1631, 2017.

[5] T. Liu, P. Maurovich-Horvat, T. Mayrhofer, et al., "Quantitative coronary plaque analysis predicts high-risk plaque morphology on coronary computed tomography angiography: results from the romicat ii trial," *The international journal of cardiovascular imaging*, vol. 34, pp. 311–319, 2018.

[6] R. C. Cury, J. Leipsic, S. Abbara, et al., "Cad-rads™ 2.0–2022 coronary artery disease-reporting and data system: an expert consensus document of the society of cardiovascular computed tomography (scct), the american college of cardiology (acc), the american college of radiology (acr), and the north america society of cardiovascular imaging (nasci)," *Cardiovascular Imaging*, vol. 15, pp. 1974–2001, 2022.

[7] C. A. Taylor, T. A. Fonte, and J. K. Min, "Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: scientific basis," *Journal of the American College of Cardiology*, vol. 61, pp. 2233–2241, 2013.

[8] A. Rossi, A. Wragg, E. Klotz, et al., "Dynamic computed tomography myocardial perfusion imaging: comparison of clinical analysis methods for the detection of vessel-specific ischemia," *Circulation: Cardiovascular Imaging*, vol. 10, p. e005505, 2017.

[9] T. Misaka, T. Furukawa, N. Asato, et al., "Perivascular fat attenuation index on non-contrast-enhanced cardiac computed tomography: comparison with coronary computed tomography angiography," *Open Journal of Radiology*, vol. 10, p. 138, 2020.

[10] Z. Sun and S. Jansen, "Personalized 3d printed coronary models in coronary stenting," *Quantitative Imaging in Medicine and Surgery*, vol. 9, p. 1356, 2019.

[11] S. Beier, J. Ormiston, M. Webster, et al., "Hemodynamics in idealized stented coronary arteries: important stent design considerations," *Annals of biomedical engineering*, vol. 44, pp. 315–329, 2016.

[12] E. E. Antoine, F. P. Cornat, and A. I. Barakat, "The stentable in vitro artery: an instrumented platform for endovascular device development and optimization," *Journal of The Royal Society Interface*, vol. 13, p. 20160834, 2016.

[13] J. N. Silva, M. Southworth, C. Raptis, and J. Silva, "Emerging applications of virtual reality in cardiovascular medicine," *JACC: Basic to Translational Science*, vol. 3, pp. 420–430, 2018.

[14] S.-J. Yoo, T. Spray, E. H. Austin III, T.-J. Yun, and G. S. van Arsdell, "Hands-on surgical training of congenital heart surgery using 3-dimensional print models," *The Journal of thoracic and cardiovascular surgery*, vol. 153, pp. 1530–1540, 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, pp. 203–211, 2021.

[17] X. Yang, L. Xu, S. Yu, Q. Xia, H. Li, and S. Zhang, "Segmentation and vascular vectorization for coronary artery by geometry-based cascaded neural network," *IEEE Transactions on Medical Imaging*, 2024.

[18] G. Zhao, K. Liang, C. Pan, et al., "Graph convolution based cross-network multiscale feature fusion for deep vessel segmentation," *IEEE transactions on medical imaging*, vol. 42, pp. 183–195, 2022.

[19] Y. Qiu, D. Shan, Y. Wang, et al., "A topology-preserving three-stage framework for fully-connected coronary artery extraction," *Medical Image Analysis*, vol. 103, p. 103578, 2025.

[20] S. Shit, J. C. Paetzold, A. Sekuboyina, et al., "cldice-a novel topology-preserving loss function for tubular structure segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 560–16 569.

[21] Y. Kirchhoff, M. R. Rokuss, S. Roy, et al., "Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures," in *European Conference on Computer Vision*. Springer, 2024, pp. 218–234.

[22] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al., "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, p. e230024, 2023.

[23] L. Mou, Y. Zhao, H. Fu, et al., "Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical image analysis*, vol. 67, p. 101874, 2021.

[24] G. Tetteh, V. Efremov, N. D. Forkert, et al., "Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes," *Frontiers in Neuroscience*, vol. 14, p. 592352, 2020.

[25] C. Metz, M. Schaap, T. van Walsum, et al., "3d segmentation in the clinic: A grand challenge ii-coronary artery tracking," *Insight Journal*, vol. 1, p. 6, 2008.

[26] H. Kirişli, M. Schaap, C. Metz, et al., "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography," *Medical image analysis*, vol. 17, pp. 859–876, 2013.

[27] R. Gharleghi, D. Adikari, K. Ellenberger, et al., "Automated segmentation of normal and diseased coronary arteries–the asoca challenge," *Computerized Medical Imaging and Graphics*, vol. 97, p. 102049, 2022.

[28] R. Gharleghi, D. Adikari, K. Ellenberger, et al., "Annotated computed tomography coronary angiogram images and associated data of normal and diseased arteries," *Scientific Data*, vol. 10, p. 128, 2023.

[29] Grand Challenge. Asoca challenge — challenge leaderboard. https://asoca.grand-challenge.org/evaluation/challenge/leaderboard/

[30] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, pp. 8766–8778, 2020.

[31] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-preserving deep image segmentation," *Advances in neural information processing systems*, vol. 32, 2019.

[32] C. Acebes, A. H. Moustafa, O. Camara, and A. Galdran, "The centerline-cross entropy loss for vessel-like structure segmentation: Better topology consistency without sacrificing accuracy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 710–720.

[33] P. Shi, J. Hu, Y. Yang, Z. Gao, W. Liu, and T. Ma, "Centerline boundary dice loss for vascular segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 46–56.

[34] M. J. Menten, J. C. Paetzold, V. A. Zimmer, et al., "A skeletonization algorithm for gradient-based optimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 394–21 403.

[35] I. Rehman and A. Rehman, "Anatomy, thorax, heart," in *StatPearls [Internet]*. StatPearls Publishing, 2023.

[36] O. A. Bamalan, F. Jozsa, and M. P. Soos, "Anatomy, thorax, heart great vessels," 2019.

[37] A. Arackal and K. Alsayouri, "Histology, heart," 2019.

[38] Cleveland Clinic. (2024) Heart: Anatomy & function. Cleveland Clinic Health Library. Last reviewed on 2024-01-26. https://my.clevelandclinic.org/health/body/21704-heart

[39] I. Ogobuiro, C. J. Wehrle, and F. Tuma, "Anatomy, thorax, heart coronary arteries," 2018.

[40] S. Kini, K. G. Bis, and L. Weaver, "Normal and variant coronary arterial and venous anatomy on high-resolution ct angiography," *American Journal of Roentgenology*, vol. 188, pp. 1665–1674, 2007.

[41] E. Halpern, *Clinical cardiac ct: Anatomy and function*, Thieme Publishers Series. Thieme, 2008. https://books.google.de/books?id=9oiWwzmYaXYC

[42] G. L. Raff, A. Abidov, S. Achenbach, et al., "Scct guidelines for the interpretation and reporting of coronary computed tomographic angiography," *Journal of cardiovascular computed tomography*, vol. 3, pp. 122–136, 2009.

[43] S. Y. Kim, J. B. Seo, K.-H. Do, et al., "Coronary artery anomalies: classification and ecg-gated multi–detector row ct findings with angiographic correlation," *Radiographics*, vol. 26, pp. 317–333, 2006.

[44] S. Jebari-Benslaiman, U. Galicia-García, A. Larrea-Sebal, et al., "Pathophysiology of atherosclerosis," *International journal of molecular sciences*, vol. 23, p. 3346, 2022.

[45] D. J. Duncker, A. Koller, D. Merkus, and J. M. Canty Jr, "Regulation of coronary blood flow in health and ischemic heart disease," *Progress in cardiovascular diseases*, vol. 57, pp. 409–422, 2015.

[46] C. Vrints, F. Andreotti, K. C. Koskinas, et al., "2024 esc guidelines for the management of chronic coronary syndromes: developed by the task force for the management of chronic coronary syndromes of the european society of cardiology (esc) endorsed by the european association for cardio-thoracic surgery (eacts)," *European heart journal*, vol. 45, pp. 3415–3537, 2024.

[47] Royal Buckinghamshire Hospital. Coronary artery disease (cad): Symptoms, causes & treatment. Health information page. https://www.royalbucks.co.uk/conditions-and-symptoms/coronary-artery-disease/

[48] J. A. Seibert, "X-ray imaging physics for nuclear medicine technologists. part 1: Basic principles of x-ray production," *Journal of nuclear medicine technology*, vol. 32, pp. 139–147, 2004.

[49] M. Berger, Q. Yang, and A. Maier, "X-ray imaging," *Medical imaging systems: an introductory guide*, pp. 119–145, 2018.

[50] Coolth and Hmilch, "Water-cooled x-ray tube (schematic)," https://commons.wikimedia.org/wiki/File:WaterCooledXrayTube.svg, 2010, public domain image from Wikimedia Commons, accessed 14 November 2025.

[51] T. Laubenberger and J. Laubenberger, *Technik der medizinischen radiologie: Diagnostik, strahlentherapie, strahlenschutz; für ärzte, medizinstudenten und mtra;[mit 71 tabellen]*. Deutscher Ärzteverlag, 1999.

[52] O. Dössel, *Bildgebende verfahren in der medizin: Von der technik zur medizinischen anwendung*, 2nd ed. Berlin, Heidelberg: Springer Vieweg, 2016. doi:10.1007/978-3-642-54407-1

[53] T. G. Mayerhöfer, S. Pahlow, and J. Popp, "The bouguer-beer-lambert law: Shining light on the obscure," *ChemPhysChem*, vol. 21, pp. 2029–2046, 2020.

[54] J. Ambrose, "Computerized transverse axial scanning (tomography): Part 2. clinical application," *The British journal of radiology*, vol. 46, pp. 1023–1047, 1973.

[55] O. Taubmann, M. Berger, M. Bögel, Y. Xia, M. Balda, and A. Maier, "Computed tomography," *Medical Imaging Systems: An Introductory Guide*, pp. 147–189, 2018.

[56] M. Mahesh and D. D. Cody, "Physics of cardiac imaging with multiple-row detector ct," *Radiographics*, vol. 27, pp. 1495–1509, 2007.

[57] J. A. Brink, J. P. Heiken, G. Wang, K. W. McEnery, F. J. Schlueter, and M. Vannier, "Helical ct: principles and technical considerations." *Radiographics*, vol. 14, pp. 887–893, 1994.

[58] P. Mah, T. Reeves, and W. McDavid, "Deriving hounsfield units using grey levels in cone beam computed tomography," *Dentomaxillofacial radiology*, vol. 39, pp. 323–335, 2010.

[59] D. Dance, S. Christofides, A. Maidment, I. McLean, and K. Ng, "Diagnostic radiology physics," *International Atomic Energy Agency*, vol. 299, pp. 12–14, 2014.

[60] T. D. DenOtter and J. Schubert, "Hounsfield unit," 2019.

[61] C. R. Becker, C. Hong, A. Knez, et al., "Optimal contrast application for cardiac 4-detector-row computed tomography," *Investigative radiology*, vol. 38, pp. 690–694, 2003.

[62] T. Flohr and B. Ohnesorge, "Cardiac gating 2," *Integrated Cardiothoracic Imaging with MDCT*, p. 23, 2010.

[63] M. Soschynski, M. T. Hagar, J. Taron, et al., "Update for the performance of ct coronary angiography–evidence-based application and technical guidance according to current consensus guidelines and practical advice from the clinical routine," in *RöFo-Fortschritte auf dem Gebiet*

der Röntgenstrahlen und der bildgebenden Verfahren, vol. 194, no. 06. Georg Thieme Verlag KG, 2022, pp. 613–624.

[64] M. Renker, U. J. Schoepf, and W. K. Kim, "Combined ct coronary artery assessment and tavi planning," *Diagnostics*, vol. 13, p. 1327, 2023.

[65] W. Birkfellner, *Applied medical image processing: a basic course*. CRC Press, 2014.

[66] R. Beare, B. Lowekamp, and Z. Yaniv, "Image segmentation, registration and characterization in r with simpleitk," *Journal of statistical software*, vol. 86, pp. 1–35, 2018.

[67] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," *Journal of digital imaging*, vol. 31, pp. 290–303, 2018.

[68] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019.

[69] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, p. 1224, 2021.

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

[71] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*. Pmlr, 2015, pp. 562–570.

[72] Python Software Foundation, "Python," https://www.python.org/, accessed: 2025-12-14.

[73] H. C. M. L. A. S.-R. H. D. S. K. J. W. R. W. T. K. A. Ewald, T. Brosch, "Omnilearn – A generalizable 3D semantic segmentation framework," Philips Research Laboratories, Hamburg, Technical Note PR-TN 2019/00232, 2019.

[74] P. A. Yushkevich, J. Piven, H. Cody Hazlett, et al., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, pp. 1116–1128, 2006.

[75] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," in *Programming with TensorFlow: solution for edge computing applications*. Springer, 2021, pp. 87–104.

[76] M. J. Cardoso, W. Li, R. Brown, et al., "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[77] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, et al., "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[78] L. Lux, A. H. Berger, A. Weers, et al., "Topograph: An efficient graph-based framework for strictly topology preserving image segmentation," *arXiv preprint arXiv:2411.03228*, 2024.

[79] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[80] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, pp. 80–83, 1945.

[81] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *International journal of endocrinology and metabolism*, vol. 10, p. 486, 2012.

[82] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[83] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[84] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific

features," in *International workshop on statistical atlases and computational models of the heart*. Springer, 2017, pp. 120–129.

[85] A. Szymczak, A. Stillman, A. Tannenbaum, and K. Mischaikow, "Coronary vessel trees from 3d imagery: a topological approach," *Medical image analysis*, vol. 10, pp. 548–559, 2006.

[86] A. Akgöz, D. Akata, T. Hazırolan, and M. Karçaaltıncaba, "Optimal reconstruction interval in dual source ct coronary angiography: a single-center experience in 285 patients," *Diagnostic and Interventional Radiology*, vol. 20, p. 399, 2014.

[87] D. Arora, N. Kumar, and S. Gupta, "Does the skeleton-recall loss really work?" *arXiv preprint arXiv:2508.11374*, 2025.

[88] A. Zeng, C. Wu, G. Lin, et al., "Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images," *Computerized Medical Imaging and Graphics*, vol. 109, p. 102287, 2023.