Geo-UNet: A Geometrically Constrained Neural Framework for Clinical-Grade Lumen Segmentation in Intravascular Ultrasound

by

Yiming Chen

S.B. Mathematics and Computer Science and Engineering,
Massachusetts Institute of Technology (2024)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2024

Authored by:	Yiming Chen Department of Electrical Engineering and Computer Science July 10, 2024
Certified by:	Polina Golland Professor of Electrical Engineering and Computer Science, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair Master of Engineering Thesis Committee

Geo-UNet: A Geometrically Constrained Neural Framework for Clinical-Grade Lumen Segmentation in Intravascular Ultrasound

by

Yiming Chen

Submitted to the Department of Electrical Engineering and Computer Science
on July 10, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Precisely estimating lumen boundaries in intravascular ultrasound (IVUS) is needed for sizing interventional stents to treat deep vein thrombosis (DVT). Unfortunately, current segmentation networks like the UNet lack the precision required for clinical adoption in IVUS workflows. This arises due to the difficulty of automatically learning accurate lumen contour from limited training data while accounting for the radial geometry of IVUS imaging. We propose the Geo-UNet framework to address these issues via a design informed by the geometry of the lumen contour segmentation task, building anatomical constraints directly into the architecture. We first convert the input data and segmentation targets from Cartesian to polar coordinates. Starting from a convUNet feature extractor, we propose a two-task setup, one for conventional pixel-wise labeling and the other for *single boundary* lumen-contour localization. We directly combine the two predictions by passing the predicted lumen contour through a new activation (named CDFeLU) to filter out spurious pixel-wise predictions. Our unified loss function carefully balances area-based, distance-based, and contour-based penalties to provide near clinical-grade generalization in unseen patient data. We also introduce a lightweight, inference-time technique to enhance segmentation smoothness. The efficacy of our framework on a venous IVUS dataset is shown against state-of-the-art models. We will make the code repository for this project available soon after approval from industry collaborators.

Thesis supervisor: Polina Golland
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

The work presented in my thesis is in collaboration with IBM research - Almaden. I am grateful to Niharika D'Souza from IBM Research for her amazing mentorship. I have learned so much from her in the past year, and I feel very lucky for the opportunity to collaborate with her. I also want to thank Dr. Tanveer Syeda-Mahmood and Dr. Kashyap Satyananda from IBM Research for their helpful feedback and kind guidance throughout the project.

I want to express my appreciation for my thesis advisor, Professor Polina Golland. Thank you so much for making me feel supported in all aspects and for warmly welcoming me into the Medical Vision Group. I have truly enjoyed the reading groups/group meetings where I gained a breadth of knowledge on medical imaging/computer vision and found a passion for this field. I want to thank Dr. Neerav Karani and Dr. Neel Dey for sharing their prior work/knowledge on this project and for answering all my questions. I also loved getting to know the rest of the lab, who are all so fun, intelligent, and vibrant. You guys are the GOATs ;).

I also want to acknowledge the collaborators at Boston Scientific for the clinical insights and feedback during my monthly update presentations.

Last but certainly not least, I want to thank my family and friends for their constant, unconditional love and support throughout this journey!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Blocked or clogged veins cause acute medical conditions with severe consequences. As a representative example, deep vein thrombosis (DVT) is a serious condition that can cause significant short-term discomfort and diminish the quality of life, potentially leading to irreversible, long-term venous system damage that may be limb or life-threatening [1]. It is a precursor to pulmonary embolism, a critical condition where a clot travels to the lungs, impeding blood oxygenation. To manage DVT, clinicians often utilize Intravascular Ultrasound (IVUS) [2] to guide endovascular treatments, where a catheter equipped with an ultrasound transducer is inserted to visualize internal structures and pinpoint anatomical landmarks.

IVUS samples are organized into pullbacks, where consecutive frames of images are captured as the catheter travels through the blood vessel. The catheter emits sound waves that are reflected by/pass through structures based on their density or acoustic impedence [2]. Dense material appears brighter. For instance, clots and fatty plaque will appear grey, and blood flow will appear black. IVUS is a radial acquisition and can be used to distinguish between key features locally around a vessel. These insights can then be used for disease-detection and treatment outcomes assessment. In the event of a clot, the physician may remove the thrombus and insert balloons or stents to keep the vessel open and ensure proper blood flow. These devices are sized based on nearby healthy regions, where accurate measurement of the vessel's lumen is crucial for avoiding complications like pain from improper device sizes or fatal stent migration [3].

Traditionally, balloon and stent-sizing can be done via manual annotation of medical images to extract the lumen. However, this is labor-intensive and not scalable. Thus, my project aims to automate this process safely and reliably by leveraging computer vision and data processing techniques. Automatic segmentation of venous IVUS (v-IVUS) images is challenging owing to variability in tissue/vessel appearance across subjects due to thin, compression-prone vessel walls, stents, artifacts, and the manual nature of the pullback (i.e. variable longitudinal frame rate across pullbacks due to manual control of the catheter by the physician).

Deep Neural Networks (DNNs) for vascular segmentation [4] have soared in popularity due to their ability to provide improved performance without manual intervention during deployment. The UNet, a fully convolutional/attentional architecture with residual connections, has shown the most success across various segmentation problems in medical imaging, even with scarce amounts of training data [5]. Variants of the UNet [6] have been successful for plaque/calcification detection and vessel segmentation [4], [7]–[9] as well as stent [10] and lesion detection/classification [11] for coronary artery disease. These use either 2D images [9] or 3D image blocks [8], [12] as inputs and produce a pixel-wise map of the segmentation target as the output. The IVUS segmentation literature focuses on arterial acquisitions which provide a different field of view (FoV) and use a motorized pullback providing a fixed longitudinal frame rate. However, venous acquisitions are not well-studied, and most existing techniques do not generalize well to v-IVUS data due to under/over-segmentation of lumen regions in the presence of imaging artifacts and their predilection to output spurious, fragmented predictions when there are nearby vessels or tissue structures. This is potentially due to their inability to reflect the radial geometry of the imaging modality and constrain the output to be a single contiguous lumen region, as dictated by the anatomy under consideration.

For my thesis, we alleviate the issues above by designing a new DNN, named Geo-UNet—a fully convolutional architecture for lumen segmentation from venous IVUS images that incorporates radial contour-geometry constraints directly in the architecture, in contrast to prior works which impose implicit anatomical constraints via loss functions [13]. Our method features 3 main components: **1) Input representation:** we operate on 2D-image inputs converted from Cartesian to polar coordinates which better reflect inherent IVUS imaging physics [9], [14]. **2) Anatomically Constrained Self-informing Network**: We propose a two-task setup with a shared UNet feature extraction module. In polar space, the lumen boundary is a single contour. While the natural prediction target is a standard pixel-level segmentation, we design a second objective to predict a single lumen boundary contour. Using this prediction as a guide, we refine the pixel-level segmentation via a new activation function—$CDFeLU$, based on the cumulative distribution function. This regularization mitigates spurious predictions from pixel-level segmentation without the need for additional post-processing, a known shortcoming of prior approaches. During training, our unified loss function combines area-based, distance-based, and contour-based penalties to improve generalization. **3) Inference-time Continuity Enhancement**: Based on the radial geometry in imaging and properties of the convolutional UNet, we propose a continuity enhancement technique, coined Geo-UNet++, which is a lightweight, inference-time procedure to address wrap-around discontinuities at $0/2\pi$ angles in the segmentation estimation. Our framework compares favorably against state-of-the-art segmentation baselines with consistent improvements in segmentation Dice scores and derived lumen diameter estimation for stent sizing.

Chapter 2

Related Works

2.1 UNet-based IVUS Segmentation Methods

The UNet has served as the foundation for feature extraction in medical segmentation tasks, and many methods build on its principles for the IVUS use case [5]. IVUS-Net [15] is a fully convolutional network with UNet as a base model and features a similar downsampling encoder and upsampling decoder framework. Within each encoder/decoder block, a “main" branch and “refining" branch separately capture features at different granularities to avoid information loss due to pooling, and their sum is the input to the next block. This model has proven effective against the task of arterial IVUS segmentation. IVUS-UNet++ [16] adds a pyramid feature network to UNet++, an enhanced UNet architecture with dense skip-connections and deep supervision on intermediate layer outputs, to consolidate feature maps at various scales. Another method combines UNet-extracted features with hand-selected features via optimization and sparse representation learning to enhance segmentation performance [17]. All of these methods take advantage of UNet's ability to capture contextual and detailed information for effective feature extraction.

2.2 Polar IVUS Representation and Relevant Smoothness-enhancement Methods

IVUS frames are captured as a catheter continuously spins and travels along the axis of longitudinal pullback. Though the images are presented with a circular Cartesian view for easy visual interpretability of the vessel, this is an inherently polar acquisition. To respect the physics of the acquisition, prior works have trained with polar space representations of IVUS samples.

Polar representations render the concentric layered geometry of Cartesian IVUS frames trivially separable as linear layers. This fact was used for more spatial context in arterial IVUS segmentation via a polar coordinate-aware network augmented with learned translation dependence [14]. In addition, a polar arterial IVUS representation was used in a Multiframe Convolutional Neural Network (MFCNN) to incorporate adjacency context [9]. In this method, to enforce a smooth prediction and a univocally defined contour position at the $0/2\pi$ warp-around, the predicted segmentation contour was assumed to be a 2π-periodic function plus Gaussian noise and is processed using a Gaussian Process regressor. Though effective in the arterial IVUS use case, the assumption of Gaussian noise on the model prediction may not hold for other models/modalities due to the variability of architectural setups and input data. The exp-sine-squared kernel used to model the periodic contour has underlying constraints on the consistency of contour positions and may not generalize to more irregular geometries. Finally, the per-frame optimization is computationally costly.

2.3 Two-Task Layer Segmentation for Retinal Optical Computed Tomography (OCT)

BoundaryReg [18] is a recent approach based on convolutional UNets designed to produce layer surface segmentation for retinal optical computed tomography (OCT). This model estimates dense pixel-wise and sparse contour predictions using a shared UNet followed by two distinct output convolution layers. It also has additional topology modules following the sparse contour predictions to ensure non-intersecting layer predictions by enforcing non-negative spacing between correctly ordered, consecutive layers. The two-task approach empirically enhances segmentation quality/stability and motivated the two-branch setup in Geo-UNet.

2.4 Anatomically Constrained Networks

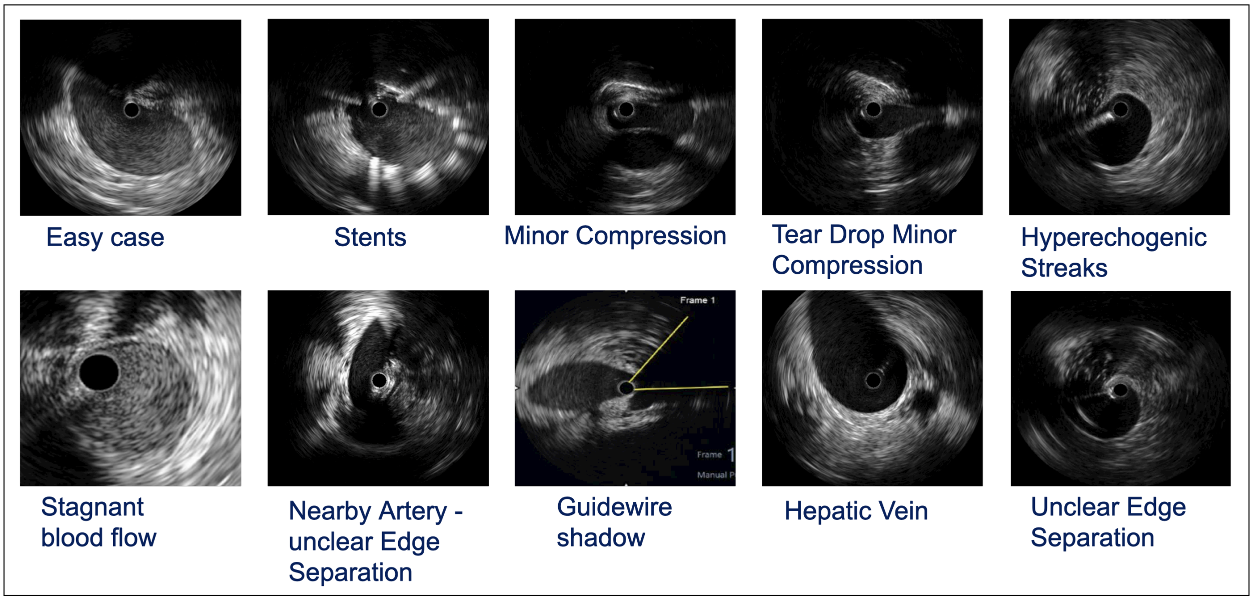
Explicit inclusion of known anatomical constraints in medical imaging tasks can help combat generalization limitations due to data scarcity or imaging artifacts. However, the incorporation of such information into CNN frameworks is not trivial. Anatomically Constrained Neural Networks (ACNNs) feature generic training strategies that places anatomical constraints via regularization [13]. Specifically, the method learns compact, non-linear relationships representative of the input anatomy via an autoencoder and encourages the model prediction to follow the learned distribution in latent space. While generalizable, this implicit regularization does not ensure correct anatomy. Modality-specific hard restrictions imposed by the network design may be a favorable choice when possible. Geo-UNet demonstrates one such instance via the contour-prediction branch guaranteeing a single-region lumen prediction—a known anatomical characteristic of a vessel lumen.

Chapter 3

Methodology

3.1 Data

3.1.1 Data Acquisition and Specifications



Easy case

Stents

Minor Compression

Tear Drop Minor Compression

Hyperechogenic Streaks

Stagnant blood flow

Nearby Artery - unclear Edge Separation

Guidewire shadow

Hepatic Vein

Unclear Edge Separation

Figure 3.1: Variations in appearance that are all considered as N1 frames with normal anatomy.

Our venous IVUS dataset is acquired using the Boston Scientific OC35 peripheral imaging catheter, which uses a rotating transducer to generate cross-sectional views. The catheter has a 70mm imaging diameter and a 15MHz operating frequency. It is typically used in the detection and treatment of venous disease (e.g. DVT, non-thrombotic iliac venous lesions, chronic post-thrombotic syndrome, and more). We obtained data for 79 patients with 166 pullbacks of varying durations. The data is labeled per frame and partitioned into two groups: diseased and normal. The former refers to regions with acute/subacute clots and chronic Post Thrombotic Syndrome (PTS). The latter contains labels N1 (frames with typical geometry despite variability in appearance shown in Figure 3.1) and N2 (frames with irregular geometry due to compression from nearby vessels but no thrombus present). Since stent-sizing is performed on healthy frames, all N1/N2 frames were labeled by expert annotators, for a total of 77,917 annotated image frames. Given the increased variability in appearance and subjectivity in annotation, the lumen in N2 frames is qualitatively harder to segment compared to N1 frames.

3.1.2 Training Data Augmentations

We apply stacked augmentations including rotation, translation, shear, contrast enhancement, Gaussian blur, intensity scaling, and speckle noise on Cartesian inputs [19]. These stacked transformations can effectively simulate the expected domain variations in a medical imaging modality, thereby enhancing model generalization during deployment.

3.2 Geo-UNet

Figure 3.2 illustrates the Geo-UNet module. Inputs and prediction targets are represented in polar space. We use a shared convolutional UNet feature extractor, connected to two distinct prediction branches via persistent skip connections and convolutional layers.

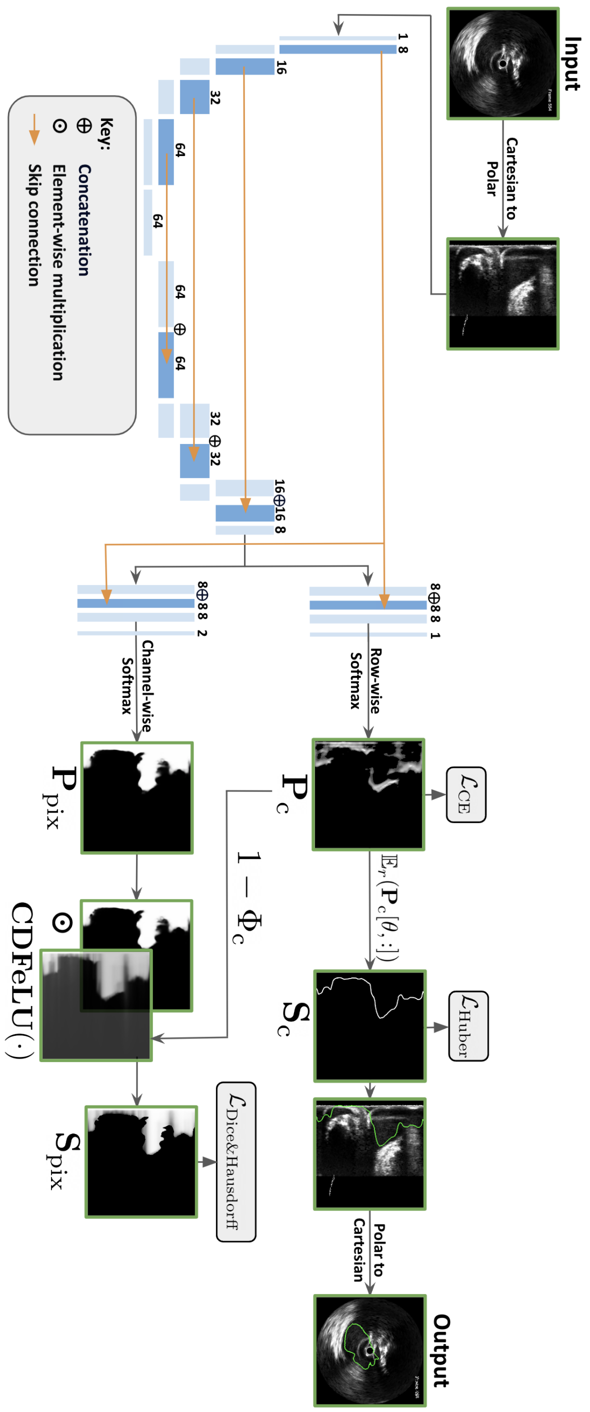

Input
Cartesian to Polar
1 8
16
32
64
64
64 ⊕ 64
32 ⊕ 32
16⊕16 8
8⊕8 8 1
8⊕8 8 2
Row-wise Softmax
Channel-wise Softmax
$\mathbf{P}_{\mathrm{c}}$
$\mathcal{L}_{\mathrm{CE}}$
$\mathbb{E}_r(\mathbf{P}_{\mathrm{c}}[\theta,:])$
$\mathbf{S}_{\mathrm{c}}$
$\mathcal{L}_{\mathrm{Huber}}$
Polar to Cartesian
Output
$1-\Phi_{\mathrm{c}}$
$\mathbf{P}_{\mathrm{pix}}$
⊙
CDFeLU(·)
$\mathbf{S}_{\mathrm{pix}}$
$\mathcal{L}_{\mathrm{Dice\&Hausdorff}}$
Key:
⊕ Concatenation
⊙ Element-wise multiplication
Skip connection

Figure 3.2: **Geo-UNet Framework for Venous Lumen Segmentation**: The feature extractor is a fully convolutional UNet module with inputs of polar 2D IVUS frames. The top branch produces a probability map for the lumen contour ($\mathbf{P}_{\mathrm{c}}$) via a row-wise softmax, which is converted to a single contour segmentation ($\mathbf{S}_{\mathrm{c}}$) via a row-wise expectation function. The bottom branch produces a per-pixel probability map ($\mathbf{P}_{\mathrm{pix}}$) via a channel-wise softmax. CDFeLU(·) allows the top branch to inform the bottom, refining the pixel-wise probabilities to give the segmentation ($\mathbf{S}_{\mathrm{pix}}$) that is compared against the (polar) ground truth lumen mask. The loss functions are highlighted in grey.

3.2.1 Cartesian to Polar Representation

As shown in Figure 3.3, the original venous IVUS dataset depicts the acquisition in Cartesian coordinates, where the catheter is a small black circle lying at the center of the square image and the field of view (FoV) is the circular region inscribed in the input image. No registration is needed to align the IVUS frames by the nature of the acquisition. We convert from x-y Cartesian space to r-θ polar space, where the horizontal axis represents the radial distance from the Cartesian origin, and the vertical axis represents the angles from 0 to 2π. This simplifies the radial geometry to a left-right geometry, where the lumen region is a vertical region on the left section of the polar image.

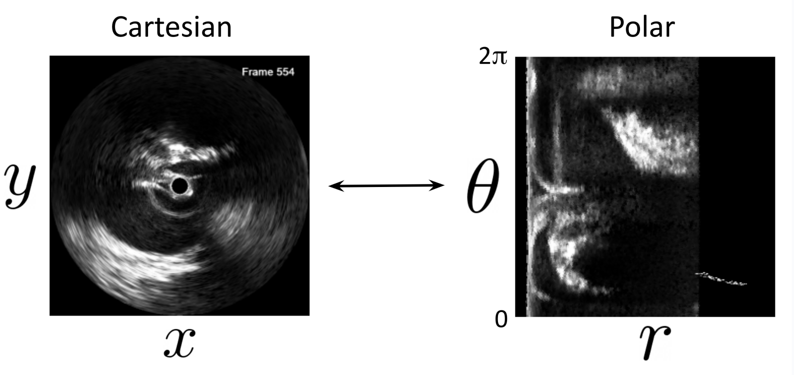


Figure 3.3: Example pair of Cartesian-polar IVUS frames.

3.2.2 Lumen Contour Estimation Branch

In polar space, the horizontal and vertical axes correspond to radii (r) and angles (θ), respectively. Let $\mathbf{Y}_{\text{pix}}$ denote the ground truth binary mask of size $R \times R$ (R=256 pixels).

To obtain the contour lumen map $\mathbf{Y}_c[\cdot]$ of size $R \times 1$, we summing along the r coordinate for each θ,

$$\mathbf{Y}_c[\theta] = \sum_r \mathbf{Y}_{\text{pix}}[\theta, r].$$

$\mathbf{Y}_c[\cdot]$ captures the lumen depth at each θ, a distinct value in {0,..., R}. The lumen boundary is a single, smooth contour with no self-intersection (i.e. has a distinct depth $r \in \{0, \dots, R-1\}$ for each $\theta \in \{0, \dots, R-1\}$, after discretizing the range $[0, 2\pi]$ into R intervals).

The top network branch captures the lumen contour by computing a softmax across each row of the single-channel output to obtain a row-sparse probability map $\mathbf{P}_{\mathrm{c}}$ of size $R \times R$. The entries $\mathbf{P}_{\mathrm{c}}[\theta, r] \in [0, 1]$ denote the probability that the contour depth at θ is r and is ideally high along the lumen contour and near 0 elsewhere. We convert $\mathbf{P}_{\mathrm{c}}$ into a segmentation contour $\mathbf{S}_{\mathrm{c}}$ with an expectation across radii values,

$$\mathbf{S}_{\mathrm{c}}[\theta] = \mathbb{E}_r(\mathbf{P}_{\mathrm{c}}[\theta, :]) = \sum_{r=0}^{R-1} r * \mathbf{P}_{\mathrm{c}}[\theta, r],$$

accounting for the uncertainty along boundary pixels in a differentiable operation [18]. This enforces a distinct contour depth for each θ.

We use two training losses for $\mathbf{P}_{\mathrm{c}}$ and $\mathbf{S}_{\mathrm{c}}$. First, we compute the cross entropy between $\mathbf{P}_{\mathrm{c}}$ and $\mathbf{Y}_c$ to promote high predicted probabilities along the contour:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{R^2} \sum_{\theta, r=0,0}^{R-1} \mathbb{1}[\mathbf{Y}_c[\theta] = r] \mathbf{log}(\mathbf{P}_{\mathrm{c}}[\theta, r]) + \mathbb{1}[\mathbf{Y}_c[\theta] \neq r] \mathbf{log}(1 - \mathbf{P}_{\mathrm{c}}[\theta, r]) \tag{3.1}$$

We then encourage the predicted segmentation $\mathbf{S}_{\mathrm{c}}[\theta]$ to match $\mathbf{Y}_c[\theta]$ using a Huber loss [20]:

$$\mathcal{L}_{\mathrm{Huber}}(\cdot) = \sum_{\theta=0}^{R-1} \frac{d_\theta^2}{2} \mathbb{1}(|\mathbf{d}_\theta| < 1) + (|\mathbf{d}_\theta| - 0.5) \mathbb{1}(|\mathbf{d}_\theta| \geq 1), \tag{3.2}$$

where $\mathbf{d}_\theta = \mathbf{Y}_c[\theta] - \mathbf{S}_{\mathrm{c}}[\theta]$. To get a polar binary segmentation that guarantees a single lumen in Cartesian space, for each θ, we have 1s for pixels to the left of/along $\mathbf{S}_{\mathrm{c}}[\theta]$ (rounded to the nearest integer) and 0 elsewhere. This yields a dense pixel mask that serves as the final prediction output.

3.2.3 Pixel-wise Segmentation Branch with Probabilistic Contour Maps

Applying a conventional channel-wise softmax operation [4], the bottom branch outputs a pixel-wise probability map $\mathbf{P}_{\text{pix}}$ of size $R \times R$, where $\mathbf{P}_{\text{pix}}[\theta, r]$ denotes the probability that pixel $[\theta, r]$ is in or on the lumen boundary. To reconcile this with the lumen contour estimate, we compute a dense probability map from $\mathbf{P}_{\text{c}}$ via a novel activation function based on the cumulative distribution function (CDF). Let $\Phi_{\text{c}}[\theta, r] = \text{CDF}(\mathbf{P}_{\text{c}}[\theta, r])$, the transformation $(1 - \Phi_{\text{c}}[\theta, r])$ models the confidence that the pixel $[\theta, r]$ is contained within the lumen and is larger at smaller radii, serving as a probabilistic mask for $\mathbf{P}_{\text{pix}}$. We compute the refined pixel-wise segmentation $\mathbf{S}_{\text{pix}}$ of size $R \times R$ by introducing the following activation,

$$\begin{aligned}\text{CDFeLU}(\mathbf{P}_{\text{pix}}, \mathbf{P}_{\text{c}}) &= \mathbf{S}_{\text{pix}}[\theta, r] \\ &= \mathbf{P}_{\text{pix}}[\theta, r] * (1 - \Phi_{\text{c}}[\theta, r]) \\ &= \mathbf{P}_{\text{pix}}[\theta, r] * \left[1 - \sum_{j=0}^{r} \mathbf{P}_{\text{c}}[\theta, j]\right].\end{aligned}$$

CDF error Linear Units (CDFeLU) is analogous to Gaussian Error Linear Units (GELU) [21], where the CDF error is estimated based on the geometry of the lumen boundary as opposed to a normal distribution. Finally, we impose a combination of area-based (Dice) and distance-based (Hausdorff [22]) losses on $\mathbf{S}_{\text{pix}}$ to match the ground truth pixel-wise lumen mask $\mathbf{Y}_{\text{pix}}$,

$$\mathcal{L}_{\text{Dice\&Hausdorff}}(\cdot) = \lambda * \mathcal{L}_{\text{Dice}}(\mathbf{S}_{\text{pix}}, \mathbf{Y}_{\text{pix}}) + (1 - \lambda) * \mathcal{L}_{\text{Haus.}}(\mathbf{S}_{\text{pix}}, \mathbf{Y}_{\text{pix}}). \tag{3.3}$$

with the trade-off $\lambda \in (0, 1)$ determined experimentally to be 0.9. Note that by design, $\text{CDFeLU}(\mathbf{P}_{\text{pix}}, \mathbf{P}_{\text{c}})$ de-emphasises regions outside the lumen (right of $\mathbf{Y}_{\mathbf{c}}[\theta]$), filtering out potentially spurious predictions in $\mathbf{P}_{\text{pix}}$, a task usually reserved for manual/semi-automated post-processing. At the same time, this achieves “communication" between the two prediction branches to reinforce overlaps between $\mathbf{P}_{\mathrm{c}}$ and $\mathbf{P}_{\mathrm{pix}}$, effectively encouraging Geo-UNet to focus on estimates that align well across the two branches during training.

Through empirical observations during training, the top contour estimation branch and the bottom pixel-wise prediction branch exhibit differences in the image features picked up as they move toward convergence. Thus, instead of merely relying on gradients flowing backward to the shared UNet feature extractor, reweighting using CDFeLU can help directly unify the distinct prediction tasks to optimize for all three losses. As illustrated in Appendix A, testing on the top contour estimation branch and the bottom pixel-wise prediction branch gave similar performances, indicating that the branches have stabilized well at convergence with our proposed technique. We take the contour prediction as the final model output simply for the single-region lumen guarantee.

3.3 Geo-UNet++: Lightweight Inference-time Continuity Enhancement

Recall that we map pixel intensities from Cartesian space to r-θ space to generate polar images, where $\theta \in \{0, \dots, 2\pi\}$. A consequence is that the intensities of the model predictions are not constrained to align at $\theta = 0$ and $\theta = 2\pi$, as they lie at the top and bottom borders of the polar image. This often results in a wrap-around discontinuity when converting back to Cartesian coordinates that consistently induces errors in the diameter estimation. To alleviate this, we introduce an inference-time technique based on the radial nature of the Cartesian v-IVUS images and properties of convolution. We apply vertical wrap-padding to yield a rectangular, continuous input ranging $\theta = \{-\pi/2, \dots, 2\pi\}$. This extension to the bottom of the polar image with an additional $\pi/2$ context near $\theta = 0$. Exploiting convolution's lack of dependence on input dimensions, we perform inference using the *same* trained Geo-UNet model. We slice the output across the middle section $\theta = \{-\pi/3, \dots, -\pi/3 + 2\pi\})$ to avoid edge effects in the padded input, before finally presenting the prediction on the original Cartesian input rotated by $\pi/3$. The rotation does not affect the clinical objective of diameter estimation from the segmentation mask. We observe improved prediction alignment along the re-sliced output for the padded input. A walk-through of the above procedures is illustrated in Figure 3.4. This technique increases deployment time marginally (0.3-0.4ms/frame) yet enhances accuracy considerably.

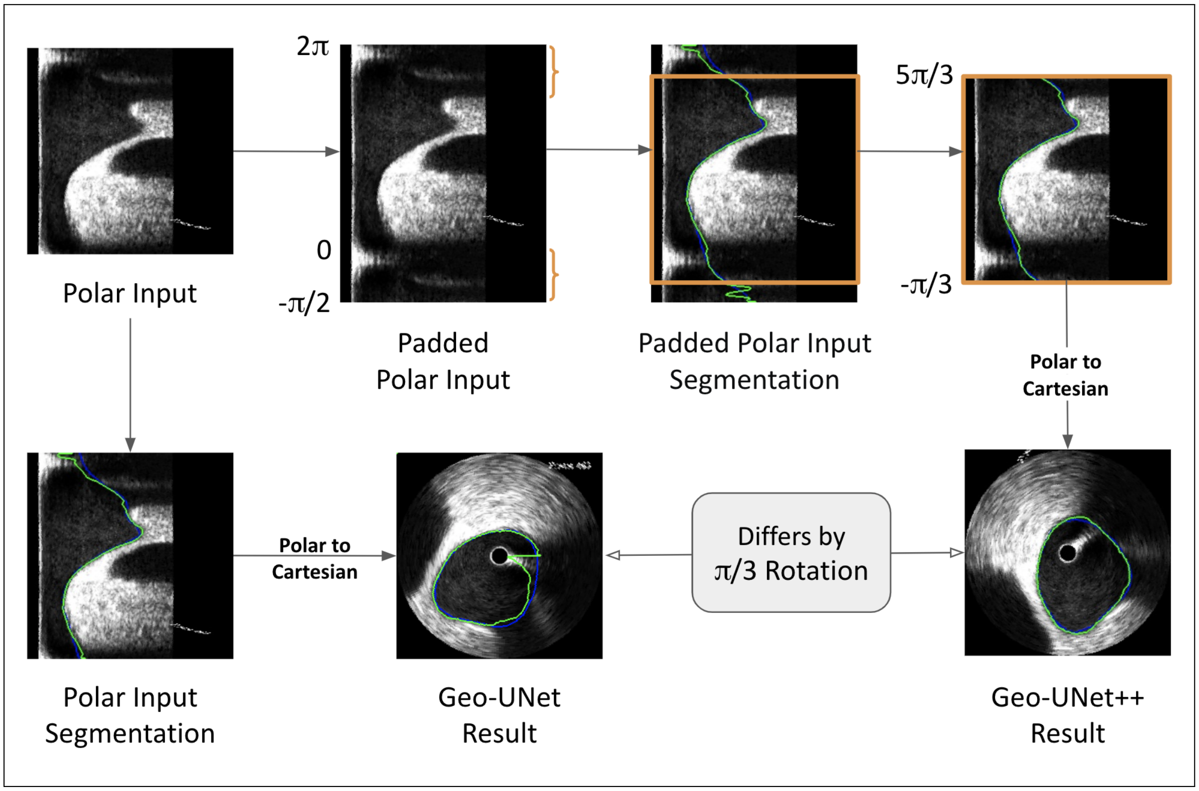


Figure 3.4: **Geo-UNet++: Inference-time Segmentation Smoothness Enhancement:** The bottom middle image shows the performance of Geo-UNet when given a polar input image. The green is the prediction, and the blue is the ground truth. In the bottom left image, note the sudden jumps and misalignment in the green prediction at the top and the bottom of the image, corresponding to 0 and 2π, respectively, yielding discontinuity at $0/2\pi$ in the Cartesian representation. Starting from the top left input image and to the right, we illustrate the ideas behind Geo-UNet++. We wrap-padded the input by copy-and-pasting the top $\pi/2$ strip to the bottom, as highlighted by the orange braces. To recover the segmentation, we take the middle portion from $\frac{-\pi}{3}$ to $\frac{5\pi}{3}$. On the lower right, we see that the Geo-UNet++ result is smoother and nearly perfectly aligned with the ground truth.

3.4 Diameter Estimation from Segmentation Mask

After obtaining the output from the top branch lumen contour prediction, we need a well-defined way to infer a diameter estimation from the segmentation mask.

3.4.1 Major and Minor Axes through the Center of Mass

One method to estimate the diameter from the segmentation mask/prediction is by passing lines through the center of mass (COM) (of the largest component) at 5° increments. The longest and shortest lengths of intersection with the mask border are the major and minor diameters, respectively. The left section of Figure 3.5 illustrates this technique, and the right sections shows the error histogram when this diameter estimation method is applied to the ground truth and the prediction of a vanilla UNet trained with only Dice loss.

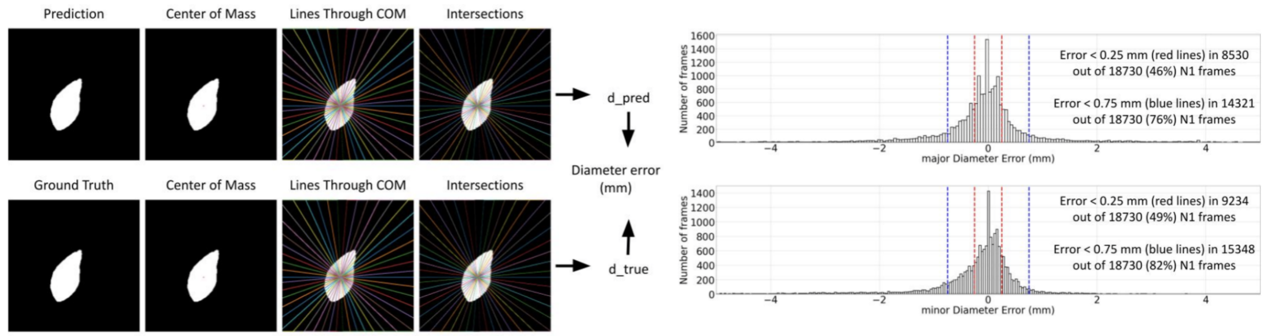


Figure 3.5: Left: Diameter estimation from the major-minor axes of COM of segmentation mask. Right: Estimation error histogram between the ground truth and predictions of a vanilla UNet trained with only Dice loss.

3.4.2 Perimeter-based Diameter Estimation

Alternatively, given the assumption of a circular vein after stent insertion, a perimeter-based diameter estimation method seemed plausible. We extract the contours of the segmentation mask (postprocessed to a single component, if needed) and divide by π to get the diameter. To mitigate the over-estimation of diameters due to concavities in the segmentation prediction, we extracted the perimeter from the mask's convex hull—the smallest convex polygon that encloses all points in the mask. However, comparing Figure 3.5 and Figure 3.6, when tested on the same vanilla UNet model trained on Dice loss only, these perimeter-based methods applied to the ground truth segmentation and the prediction showed greater errors. As a result, we resort to the major/minor axes of COM method in evaluating the performance of Geo-UNet.

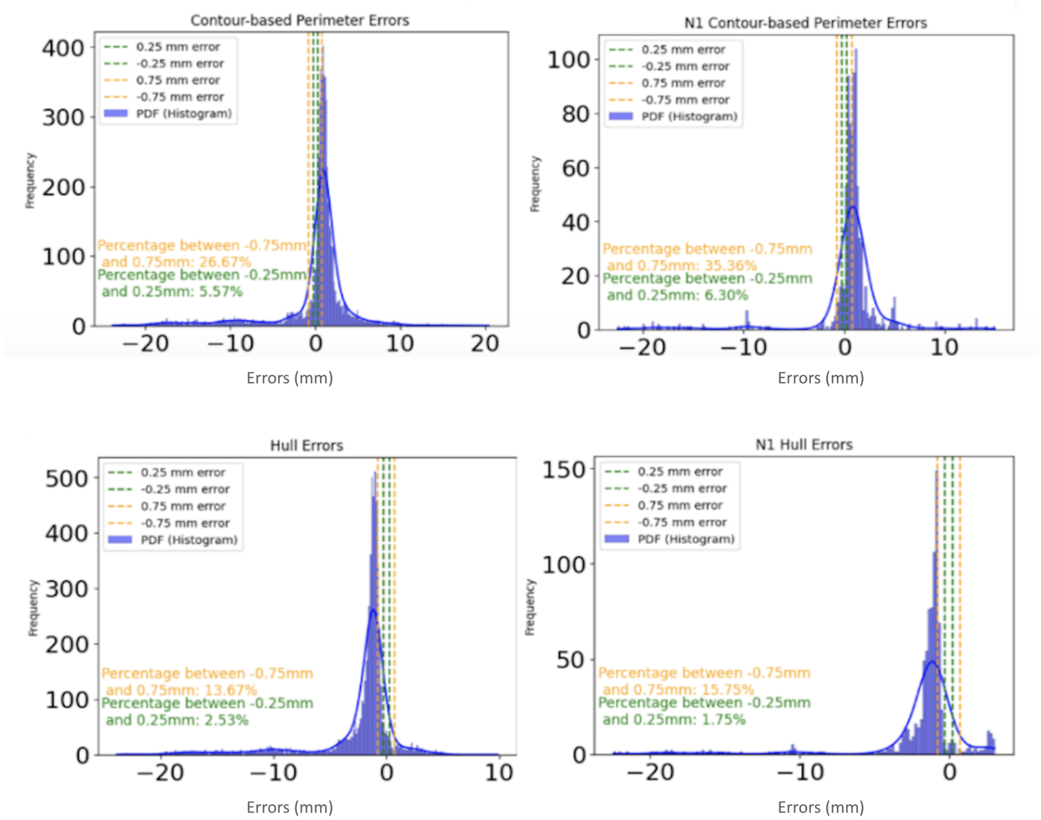


Figure 3.6: Various Perimeter-based Diameter Estimation Results.

Chapter 4

Results

4.1 Baseline Comparisons

We curate our baselines to reflect the state-of-the-art in the fields of medical image segmentation and automated processing of IVUS images.

4.1.1 MedSAM

Medical Segment Anything Model [23] is a general-purpose, promptable 2D-segmentation model with a ViT backbone [24], trained on multiple modalities (CT, MRI, ultrasound, etc). The inputs are 2D medical images and a user-specified bounding box to produce a binary pixel-wise segmentation without fine-tuning. We input the Cartesian v-IVUS images and a fixed bounding box based on the FoV to accommodate lumen regions with the largest diameters.

4.1.2 BoundaryReg

Geo-UNet's two-task prediction objective was inspired by BoundaryReg for OCT layer segmentation [18]. As A-scan OCT images and retinal layer segmentation have analogous geometric properties to polar v-IVUS representations and lumen boundary estimation, respectively, we implement this baseline for our application according to the architectural details presented in the paper (we used the same loss functions as those of Geo-UNet, which are better tailored toward IVUS and empirically verified to show better performance.)

4.1.3 Cartesian Dice + Hausdorff

Convolutional UNets are commonly used for lumen segmentation from 2D (arterial) IVUS images [4]. To adopt these baselines to v-IVUS, we use the architecture from Figure 3.2 with only the bottom branch where inputs are Cartesian v-IVUS images and outputs are Cartesian masks. We train using $\mathcal{L}_{\text{Dice\&Hausdorff}}(\cdot)$ between predictions and ground truths [22], [25].

4.1.4 Polar Dice + Hausdorff

In line with prior work [9], [14], we adopt a similar architecture and loss function as the previous baseline, but convert the inputs and targets to polar representations. This baseline also serves as an ablation for Geo-UNet where the contribution of the contour estimation branch is omitted. We obtain a single lumen region from the potentially fragmented pixel-wise predictions by post-processing the outputs to retain the largest connected component [4], both in this approach and the previous baseline.

4.2 Ablation Studies

To evaluate Geo-UNet, we perform two ablations that systematically remove its key constituent components. These comparisons are (1) Geo-UNet excluding the CDFeLU reweighting and (2) Geo-UNet without the pixel-wise prediction branch. The former uses the same loss function as Geo-UNet while the latter trains the model on a combination of $\mathcal{L}_{\text{CE}}(\cdot)$ and $\mathcal{L}_{\text{Huber}}(\cdot)$.

4.3 Implementation Details

We train all models on healthy images (frames marked N1 and N2) and adopt a three-fold cross-validation which stratifies pullbacks across patients (53/21/5 train/test/validation). Augmented input IVUS frames and model outputs are of size 256 × 256 ($R = 256$). Hyperparameters across all experiments are determined using the validation set. We use a batch size of 3 with 16 gradient accumulation steps. The Adam optimizer is used with a scheduler that linearly decreases the learning rate from 10^{-4} to 10^{-7} over 50,000 training iterations.

We use the validation set to update the weights of the best exponential moving average (EMA) model. To save on compute time, we only retain $\mathcal{L}_{\text{Huber}}(\cdot)$ (Eq.(3.2)) at each validation step to guide optimization for Geo-UNet, BoundaryReg, and the first ablation study removing CDFeLU from Geo-UNet. Similarly, we kept only $\mathcal{L}_{\text{Dice}}(\cdot)$ for the Cartesian and polar UNet baselines. Our machine has 50 CPU cores and 2 A-100 NVIDIA GPUs with 32GB RAM, resulting in an average training time of 3.5-4 hrs per cross-validation fold.

4.4 Clinical Targets

In addition to the Dice score at test time, we evaluate the measurement error in the diameter of the major/minor axes of the predicted lumen against that of the ground truth lumen [3]. Commercial stents are sized on N1 frames, are available in 0.5mm increments, and are sized against the average of the major and minor diameter [3]. Per a clinician, the models should ideally achieve a major and minor axis diameter error within 0.25/0.5/0.75mm for 50/90/95% of all N1 frames. N2 frames are mainly used for vessel compression detection and not for stent-sizing. Thus, they have less stringent clinical targets of 50/70% of frames within errors of 0.5/0.75mm.

4.5 Lumen Segmentation Performance Analysis

To quantify the generalization performance, we report the test-Dice and percentage of frames with major and minor diameter error within 0.25/0.5/0.75mm for N1 and N2 frames in the test subjects in Table 4.1 for all models.

Despite being trained on ultrasound modalities, we observe that MedSAM [23] severely under-performs all the conv-UNet frameworks trained on v-IVUS, due to generalization limitations and an inability to meaningfully discern the lumen region without a more carefully curated manual prompt. BoundaryReg underperforms Geo-UNet due to architectural differences and the lack of IVUS anatomy-rooted design decisions. The model trained in Cartesian space uniformly performs worse than all polar models, reinforcing our choice to use polar representations. The polar UNet trained on only pixel-wise segmentation performs worse than Geo-UNet on several comparisons. Upon a qualitative examination (see fourth row of Figure 4.1), the last two baselines can result in fragmented predictions with multiple components, as they are not constrained to predict a single lumen contour. This problem is not resolved by post-processing to choose the largest component given the heterogeneity across pullbacks and anatomical locations. Taking the output from the contour prediction branch inherently ensures a single prediction region. The combination of the two branches is effective as seen by comparing Geo-UNet and its ablated version without the pixel-wise prediction. Removing the re-weighting (CDFeLU) worsens performance on both N1 and N2 frames. Finally, Geo-UNet++, featuring continuity enhancement during inference, provides improvements in the estimates of the minor diameter, while maintaining the quality of the major diameter estimates for the N1 frames [1]. Overall, these observations make a strong case for adopting geometry-informed principles into the design of neural frameworks for lumen segmentation from v-IVUS imaging.

[1] Errors on N2 major diameters remain above clinical precision despite slightly worsening

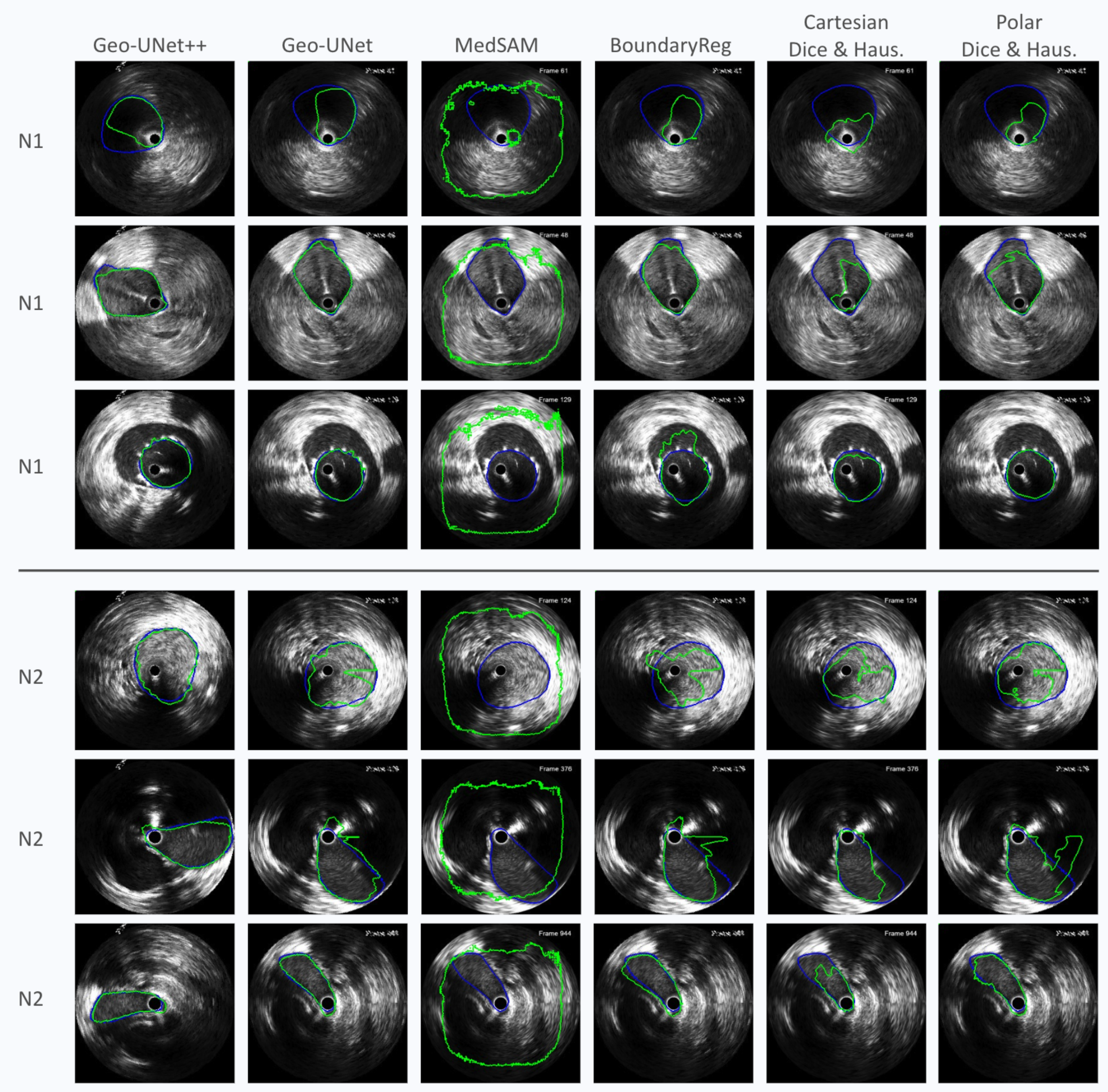


Figure 4.1: Example lumen segmentation performance of Geo-UNet++, Geo-UNet, and baselines. Note that Geo-UNet++ shows the segmentation on a rotated Cartesian input by the nature of the method. On the third row, we see a stented N1 frame. The ground truth characterizes the stented region as the lumen though there appears to be a distinctive surrounding vessel border; most methods (including Geo-UNet) performed well in this case.

Table 4.1: Lumen Segmentation performance of Geo-UNet, baselines and ablations. The best performance is in bold, while the second to best is underlined.

Methodology	Test Dice (avg/stddev)	% Frames w. Major Dia. within 0.25/0.50/0.75mm	% Frames w. Minor Dia. within 0.25/0.50/0.75mm
	Against Baselines (N1 frames)		
Geo-UNet++	0.95/0.045	66/**84**/**90**	**73**/**89**/**94**
Geo-UNet	**0.95/0.034**	**69**/**84**/**90**	69/85/91
MedSAM [23]	0.31/0.087	0/0/0	0/0/0
BoundaryReg [18]	0.94/0.043	60/78/86	70/86/91
Cartesian Dice + Haus.	0.93/0.051	61/77/83	62/79/87
Polar Dice + Haus.	0.94/0.038	66/80/87	67/84/90
	Against Baselines (N2 frames)		
Geo-UNet++	**0.88/0.094**	41/59/69	**60**/**80**/**87**
Geo-UNet	0.87/0.10	**47**/**64**/**73**	57/76/85
MedSAM [23]	0.23/0.085	0/0/0	0/0/0
BoundaryReg [18]	0.87/0.093	36/54/65	55/74/84
Cartesian Dice + Haus.	0.83/0.12	32/44/52	44/63/74
Polar Dice + Haus.	0.86/0.12	40/58/69	55/74/83
	Against Ablations (N1 frames)		
Geo-UNet	**0.95/0.034**	**69**/**84**/**90**	**69**/**85**/**91**
w/o CDFeLU reweight.	0.94/0.035	**69**/82/88	65/83/90
w/o pixel-wise pred.	0.95/0.039	67/81/87	**69**/**85**/**91**
	Against Ablations (N2 frames)		
Geo-UNet	0.87/0.10	**47**/**64**/**73**	**57**/**76**/**85**
w/o CDFeLU reweight.	0.86/0.10	45/63/72	53/71/81
w/o pixel-wise pred.	**0.88/0.092**	46/62/71	**57**/**76**/**85**

Chapter 5

Other Attempted Enhancements

Prior to finalizing the Geo-UNet framework, we explored other modifications in an effort to increase the segmentation performance. Attempts involved architectural changes and additional regularization. Though we saw either decreased performance or no significant improvements with the following additions, the lessons learned from empirical explorations were instructional.

5.1 UNet Feature Extractor Sharing Levels

The UNet excels at the medical image segmentation task partially due to the long-range skip connections. By directly concatenating earlier layers to the corresponding layers in the decoder path, the skip connections preserve high spatial resolution details lost during down sampling. They allow the network to combine high-level features from deeper layers with the low-level features from earlier layers to precisely locate pixels near segmentation borders.

With the two-task design, we were curious about how much of the UNet feature extractor should be shared by the two branches. In other words, we wondered how much weight customization each task should receive after the UNet diverges into separate prediction branches to optimize performance. Given the difficulty of interpreting intermediate latent layers, we trained three models with varying levels of UNet feature extractor sharing using the losses from Geo-UNet without CDFeLU, the architectures are shown in Figure 5.1.

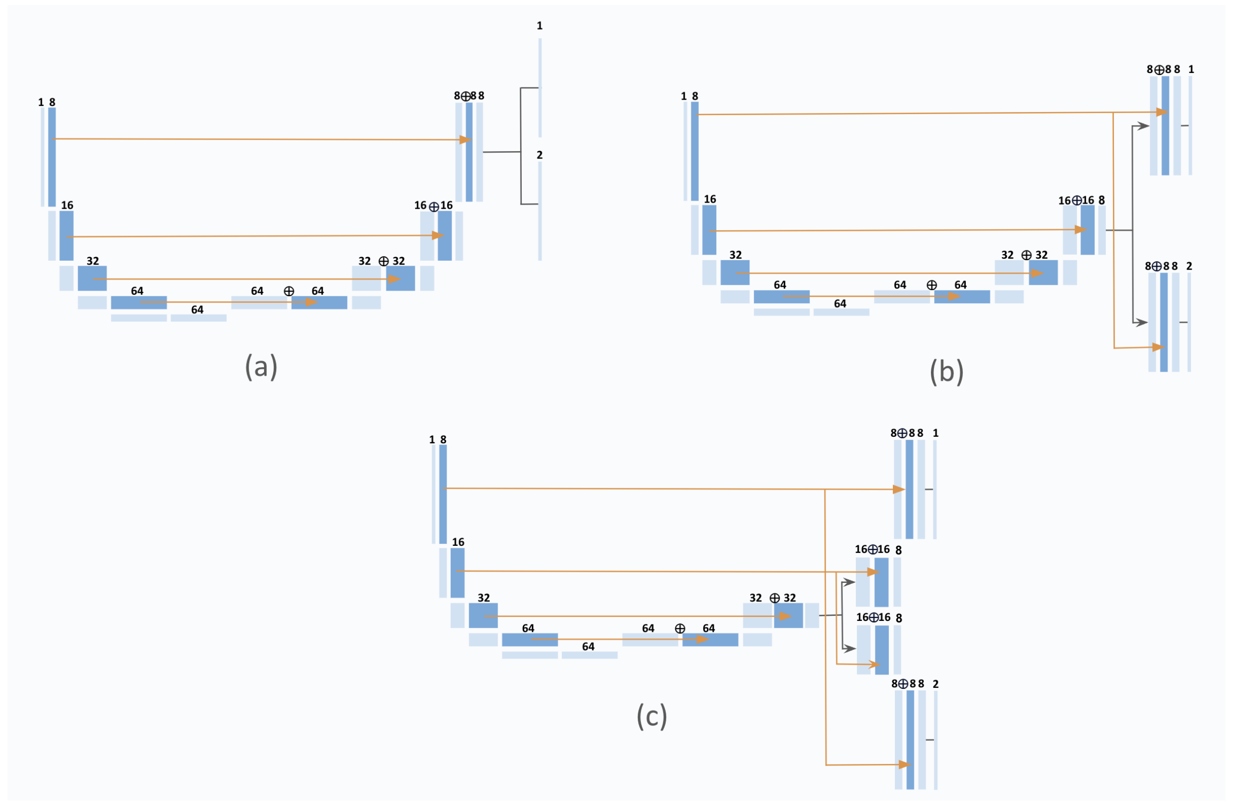


Figure 5.1: Two task prediction architecture with varying UNet sharing levels. (a) All of the encoder and decoder path is shared until the final output convolution. Same architecture and experiment as BoundaryReg. (b) Sharing ends before the final skip connection. Same architecture as Geo-UNet. (c) Sharing ends before the penultimate skip connection.

As shown in Table 5.1, sharing the UNet until the last skip-connections consistently outperforms the others in major diameter estimation and is comparable in terms of test Dice average and standard deviation. Although it is slightly worse on the minor diameter estimation, we settled on this architectural component for Geo-UNet weighing its overall performance and network simplicity.

5.2 Pixel-to-Contour Regularization

CDFeLU($\mathbf{P}_{\text{pix}}, \mathbf{P}_{\text{c}}$) combines the two prediction objectives by using the contour prediction to reweight the pixel-wise prediction. In order to allow explicit “two-way" communication

Table 5.1: Lumen segmentation performance on two task prediction model (Geo-UNet without CDFeLU) with varying UNet Feature Extractor Sharing Levels. Best result is in bold; second best result is underlined.

Shared UNet until ...	Test Dice (avg/stddev)	% Frames w. Major Dia. within 0.25/0.50/0.75mm	% Frames w. Minor Dia. within 0.25/0.50/0.75mm
		N1 frames	
(a) Last conv. layer	0.94/0.043	60/78/86	**70/86/91**
(b) Last skip-connection	0.94/0.035	**69/82/88**	65/83/90
(c) Penultimate skip-connection	**0.95/0.044**	65/81/87	**70/86/91**
		N2 frames	
(a) Last conv. layer	**0.87/0.093**	36/54/65	55/74/84
(b) Last skip-connection	0.86/0.102	**45/63/72**	53/71/81
(c) Penultimate skip-connection	**0.87/0.10**	39/57/69	**57/77/86**

between the prediction branches, we investigated the following regularization where the pixel-wise prediction informs the contour prediction.

Across each row in the dense pixel-wise predicted probabilities, an accurate segmentation will have near 1s for pixels until the border, and near 0s afterward. Thus, summing up all the radii values across each row should give a contour depth that is representative of the contour depicted by the pixel-wise prediction. As a result, we proposed an additional $\mathcal{L}_{\text{Huber}}(\cdot)$ between the $\mathbf{P}_{\text{c}}$ and the contour $\mathbf{C}_{\text{pix}}[\theta]$ derived from $\mathbf{P}_{\text{pix}}$, where

$$\mathbf{C}_{\text{pix}}[\theta] = \sum_{r=0}^{R-1} \mathbf{P}_{\text{pix}}[\theta, r].$$

Since the intuition for this regularization is based on a reasonable pixel-wise prediction, we introduce it as a fine-tuning mechanism on a Geo-UNet that is 1) trained for 50,000 iterations

Table 5.2: Segmentation performance with Pixel-to-Contour Regularization. Best result is in bold; second best result is underlined.

Method	Test Dice (avg/stddev)	% Frames w. Major Dia. within 0.25/0.50/0.75mm	% Frames w. Minor Dia. within 0.25/0.50/0.75mm
		N1 frames	
Geo-UNet	0.95/0.034	**69**/**84**/**90**	69/85/91
Finetune from 50,000 iters	0.95/0.037	66/83/89	**71**/**88**/**92**
Finetune from 10,000 iters	**0.95/0.030**	64/81/88	**71**/87/**92**
		N2 frames	
Geo-UNet	0.87/0.10	**47**/**64**/**73**	57/76/85
Finetune from 50,000 iters	**0.88/0.11**	42/62/70	**61**/**81**/**89**
Finetune from 10,000 iters	0.87/0.10	40/58/69	59/78/86

and 2) trained for 10,000 iterations, when convergence begins to taper off. Fine-tuning is done until model convergence after introduction of the new regularization.

However, as shown in Table 5.2, fine-tuning with this regularization term did not yield a meaningful improvement in segmentation performance across all metrics and thereby was not included in Geo-UNet.

5.3 Segmentation Boundary Smoothing

Given the anatomical constraints, lumen boundaries exhibit local smoothness. While the segmentation prediction from Geo-UNet is already smooth on frames for which it is accurate, we hypothesized that explicit smoothing using a 1D average pooling layer across the vertical axis/θ values of the predicted probabilities on both branches helps enforce this assumption. Specially, Figure 5.2 is a visual depiction of the effect of an average pooling procedure on the contour and pixel-wise predicted probabilities for kernel sizes of 5 and 11. We incorporated pooling layers with varying kernel sizes during training, a subset of which are shown in 5.3. Although a kernel size of 3 gave significant improvements for the major diameter estimation of N2 frames, it did not yield a meaningful improvement for the more clinically relevant N1 frames. Moreover, the pooling layer adds a substantial amount of training time. Incorporating it is an unworthy trade-off between performance and speed.

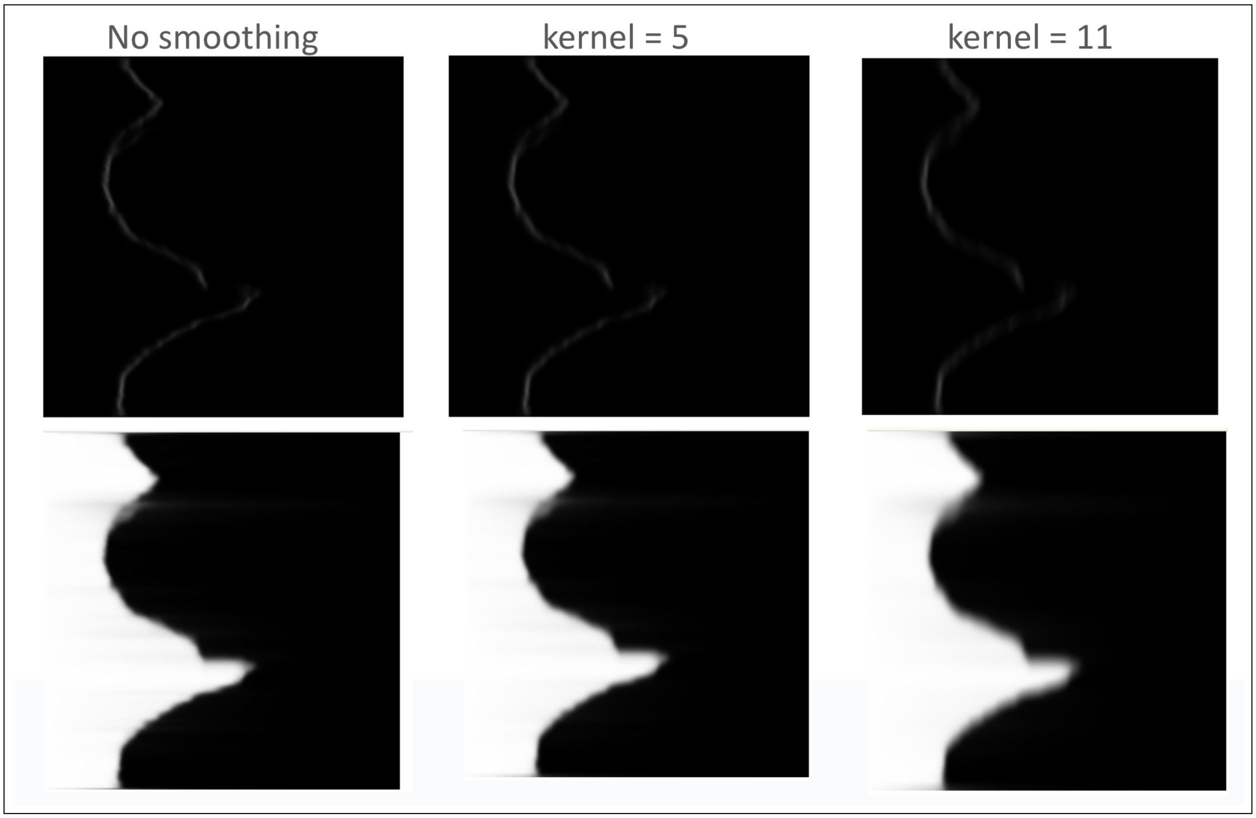


Figure 5.2: 1D segmentation boundary smoothing for kernel sizes of 5 and 11.

Table 5.3: Segmentation performance with 1d average pooling of varying kernel sizes. Best result is in bold; second best result is underlined.

Kernel size	Test Dice (avg/stddev)	% Frames w. Major Dia. within 0.25/0.50/0.75mm	% Frames w. Minor Dia. within 0.25/0.50/0.75mm
	N1 frames		
0 (**Geo-UNet**)	**0.95/0.034**	69/84/90	69/85/91
3	0.94/0.044	**71/85/91**	69/85/91
5	0.95/0.040	67/82/88	**70/86/92**
	N2 frames		
0 (**Geo-UNet**)	0.87/0.10	47/64/73	57/76/85
3	0.87/0.10	**58/77/86**	56/75/85
5	**0.88/0.093**	46/65/74	**58/77/86**

Chapter 6

Discussion and Conclusions

We develop a novel geometry-informed neural model, Geo-UNet, for precise lumen segmentation on venous IVUS imaging for automated stent-sizing. The two-task design, i.e. lumen contour estimation and dense pixel prediction, ensures appropriate constraints per data geometry and enables the self-informing feature of the two branches, thereby maximizing the capabilities of a shared UNet feature extractor. The CDFeLU re-weighting allows us to unify the distinct prediction targets probabilistically and effectively mitigate spurious predictions. The inclusion of complementary losses for each prediction target provides sufficient regularization to ensure reliable and robust generalization across unseen patients and pullbacks despite the modest dataset size. Finally, the inference time enhancement takes advantage of the input transformation and improves performance with negligible cost. Overall, Geo-UNet/Geo-UNet++ achieves a majority of clinical targets, with only a narrow gap in others, making it an attractive assistive tool for interventional specialists.

6.1 Future Work

Having optimized for architecture design and loss/regularization formulations, we believe that the Geo-UNet model is nearing a performance ceiling for 2D segmentation frameworks on venous IVUS data. However, one avenue for performance enhancement is incorporating temporal information from consecutive frames to enforce segmentation smoothness, essentially extending the problem to a 3D segmentation task. Prior works employ techniques for temporal context and smoothness on arterial IVUS data [14] .

Applying this method directly to v-IVUS is potentially challenging for three reasons. 1) Due to thinner vessel walls, venous IVUS can take on a wide range of shapes. 2) The N1/N2 training frames often constitute non-contiguous sections within a single pullback due to the presence of interspersed and anatomically distinct diseased frames. Thus, the incorporation of temporal information must be sectional and selective. 3) The variable frame rates due to the manual nature of v-IVUS pullbacks must be deliberated, perhaps with local constant rate assumptions. Taking inspiration from a recent work titled NeuralOCT which utilizes point cloud extraction from 2D segmentation masks to achieve neural 3D reconstruction of airway OCT geometry [26], we hope to stitch together individual Geo-UNet prediction masks with appropriate parameterization to promote temporal continuity.

Appendix A

Test-time Performance Comparison across the Two Branches in Geo-UNet

Recall that the model output is given by the sparse contour prediction branch due to its guarantee of a single-connected lumen. However, as shown in Table A.1, for the same trained Geo-UNet model, the two branches yield comparable performance across all metrics. This indicates that Geo-UNet is desirably consistent and stable across different prediction objectives.

Table A.1: Test-time performance comparison across the sparse contour prediction branch and the dense pixel-wise prediction branch in Geo-UNet

Prediction branch	Test Dice (avg/stddev)	% Frames w. Major Dia. within 0.25/0.50/0.75mm	% Frames w. Minor Dia. within 0.25/0.50/0.75mm
		N1 frames	
Contour	0.95/0.034	69/84/90	69/85/91
Pixel-wise	0.94/0.040	67/82/89	69/85/91
		N2 frames	
Contour	0.87/0.10	47/64/73	57/76/85
Pixel-wise	0.88/0.099	42/59/68	56/75/84

References

[1] D. Scarvelis and P. S. Wells, "Diagnosis and treatment of deep-vein thrombosis," *Cmaj*, vol. 175, no. 9, pp. 1087–1092, 2006.

[2] E. A. Secemsky, S. A. Parikh, M. Kohi, M. Lichtenberg, M. Meissner, R. Varcoe, A. Holden, M. Jaff, D. Chalyan, D. Clair, *et al.*, "Intravascular ultrasound guidance for lower extremity arterial and venous interventions," *EuroIntervention*, vol. 18, no. 7, p. 598, 2022.

[3] P. Stähr, H.-J. Rupprecht, T. Voigtländer, P. Kearney, R. Erbel, L. Koch, S. Kraß, R. Brennecke, and J. Meyer, "Importance of calibration for diameter and area determination by intravascular ultrasound," *The International Journal of Cardiac Imaging*, vol. 12, pp. 221–229, 1996.

[4] P. Arora, P. Singh, A. Girdhar, and R. Vijayvergiya, "A state-of-the-art review on coronary artery border segmentation algorithms for intravascular ultrasound (IVUS) images," *Cardiovascular Engineering and Technology*, vol. 14, no. 2, pp. 264–295, 2023.

[5] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–*

MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.

[7] M. Xie, Y. Li, Y. Xue, L. Huntress, W. Beckerman, S. A. Rahimi, J. W. Ady, and U. W. Roshan, "Two-stage and dual-decoder convolutional U-Net ensembles for reliable vessel and plaque segmentation in carotid ultrasound images," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2020, pp. 1376–1381.

[8] X. Huang, R. Bajaj, Y. Li, X. Ye, J. Lin, F. Pugliese, A. Ramasamy, Y. Gu, Y. Wang, R. Torii, *et al.*, "Post-IVUS: A perceptual organisation-aware selective transformer framework for intravascular ultrasound segmentation," *Medical Image Analysis*, vol. 89, p. 102 922, 2023.

[9] P. J. Blanco, P. G. Ziemer, C. A. Bulant, Y. Ueki, R. Bass, L. Räber, P. A. Lemos, and H. M. Garcia-Garcia, "Fully automated lumen and vessel contour segmentation in intravascular ultrasound datasets," *Medical image analysis*, vol. 75, p. 102 262, 2022.

[10] T. Wissel, K. A. Riedl, K. Schaefers, H. Nickisch, F. J. Brunner, N. D. Schnellbaecher, S. Blankenberg, M. Seiffert, and M. Grass, "Cascaded learning in intravascular ultrasound: Coronary stent delineation in manual pullbacks," *Journal of Medical Imaging*, vol. 9, no. 2, pp. 025 001–025 001, 2022.

[11] L. Meng, M. Jiang, C. Zhang, and J. Zhang, "Deep learning segmentation, classification, and risk prediction of complex vascular lesions on intravascular ultrasound images," *Biomedical Signal Processing and Control*, vol. 82, p. 104 584, 2023.

[12] S. Kashyap, N. Karani, A. Shang, N. D'Souza, N. Dey, L. Jain, R. Wang, H. Akakin, Q. Li, W. Li, *et al.*, "Feature selection for malapposition detection in intravascular ultrasound-a comparative study," in *International Workshop on Applications of Medical AI*, Springer, 2023, pp. 165–175.

[13] O. Oktay, E. Ferrante, K. Kamnitsas, *et al.*, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2018. DOI: 10.1109/TMI.2017.2743464.

[14] M. Szarski and S. Chauhan, "Improved real-time segmentation of intravascular ultrasound images using coordinate-aware fully convolutional networks," *Computerized medical imaging and graphics*, vol. 91, p. 101 955, 2021.

[15] J. Yang, L. Tong, M. Faraji, and A. Basu, *IVUS-Net: An intravascular ultrasound segmentation network*, 2018. arXiv: 1806.03583 [stat.ML]. URL: https://arxiv.org/abs/1806.03583.

[16] F. Zhu, Z. Gao, C. Zhao, *et al.*, "A deep learning-based method to extract lumen and media-adventitia in intravascular ultrasound images," *Ultrasonic Imaging*, 2022. DOI: 10.1177/01617346221114137. URL: https://doi.org/10.1177/01617346221114137.

[17] K. Li, J. Tong, X. Zhu, and S. Xia, "Automatic lumen border detection in IVUS images using deep learning model and handcrafted features," *Ultrasonic Imaging*, vol. 43, pp. 59–73, 2021. DOI: 10.1177/0161734620987288.

[18] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Fully convolutional boundary regression for retina oct segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 11764, pp. 120–128, Oct. 2019, Epub 2019 Oct 10. DOI: 10.1007/978-3-030-32239-7_14.

[19] L. Zhang, X. Wang, D. Yang, *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020, Epub 2020 Feb 12. DOI: 10.1109/TMI.2020.2973595.

[20] P. J. Huber, "Robust estimation of a location parameter," *Annals of Statistics*, vol. 53, no. 1, pp. 73–101, 1964. DOI: 10.1214/aoms/1177703732.

[21] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[22] M. Cardoso, W. Li, R. Brown, and et al., "MONAI: An open-source framework for deep learning in healthcare," *arXiv*, 2022, Published online November 4, 2022. URL: https://arxiv.org/abs/2211.02701.

[23] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, Jan. 2024, ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z. URL: http://dx.doi.org/10.1038/s41467-024-44824-z.

[24] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomedical Signal Processing and Control*, vol. 84, p. 104 791, 2023.

[25] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[26] Y. Jiao, A. Oldenburg, Y. Xu, S. Soundararajan, C. Zdanski, J. Kimbell, and M. Niethammer, *NeuralOCT: Airway oct analysis via neural fields*, 2024. arXiv: 2403.10622 [eess.IV]. URL: https://arxiv.org/abs/2403.10622.