POST-IVUS: a perceptual organisation-aware selective transformer framework for intravascular ultrasound segmentation

Huang, X.R.; Bajaj, R.; Li, Y.L.; Ye, X.; Lin, J.; Pugliese, F.; ... ; Zhang, Q.N.

Medical Image Analysis 89 (2023) 102922

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

POST-IVUS: A perceptual organisation-aware selective transformer framework for intravascular ultrasound segmentation

Xingru Huang [a,d,2], Retesh Bajaj [b,c,2], Yilong Li [a], Xin Ye [e], Ji Lin [a], Francesca Pugliese [b,c], Anantharaman Ramasamy [b,c], Yue Gu [f], Yaqi Wang [g], Ryo Torii [h], Jouke Dijkstra [i], Huiyu Zhou [j], Christos V. Bourantas [b,c], Qianni Zhang [a,*]

[a] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E3 4BL, UK
[b] Department of Cardiology, Barts Heart Centre, Barts Health NHS Trust, West Smithfield, London, EC1A 7BE, UK
[c] Centre for Cardiovascular Medicine and Devices, William Harvey Research Institute, Queen Mary University of London, London, UK
[d] School of Communication Engineering, Hangzhou Dianzi University, Xiasha Higher Education Zone, Hangzhou, Zhejiang, China
[e] Zhejiang Provincial People's Hospital, 270 West Xueyuan Road, Wenzhou, Zhejiang, China
[f] Zhejiang Institute of Mechanical and Electrical Engineering, Hangzhou, China
[g] College of Media Engineering, Communication University of Zhejiang, Hangzhou, China
[h] Department of Mechanical Engineering, University College London, London, UK
[i] Leiden University Medical Center, Leiden, Netherlands
[j] School of Informatics, University of Leicester, University Road, Leicester, LE1 7RH, United Kingdom

ARTICLE INFO

Keywords:
Intravascular ultrasound
Atherosclerosis
Semantic segmentation
Plaque burden

ABSTRACT

Intravascular ultrasound (IVUS) is recommended in guiding coronary intervention. The segmentation of coronary lumen and external elastic membrane (EEM) borders in IVUS images is a key step, but the manual process is time-consuming and error-prone, and suffers from inter-observer variability. In this paper, we propose a novel perceptual organisation-aware selective transformer framework that can achieve accurate and robust segmentation of the vessel walls in IVUS images. In this framework, temporal context-based feature encoders extract efficient motion features of vessels. Then, a perceptual organisation-aware selective transformer module is proposed to extract accurate boundary information, supervised by a dedicated boundary loss. The obtained EEM and lumen segmentation results will be fused in a temporal constraining and fusion module, to determine the most likely correct boundaries with robustness to morphology. Our proposed methods are extensively evaluated in non-selected IVUS sequences, including normal, bifurcated, and calcified vessels with shadow artifacts. The results show that the proposed methods outperform the state-of-the-art, with a Jaccard measure of 0.92 for lumen and 0.94 for EEM on the IVUS 2011 open challenge dataset. This work has been integrated into a software QCU-CMS[1] to automatically segment IVUS images in a user-friendly environment.

1. Introduction

According to the World Health Organization (WHO), cardiovascular disease is the number one cause of death worldwide (Kaptoge et al., 2019). Statistics show that about 18.6 million people died of cardiovascular disease globally in 2019 (Roth et al., 2020). Symptoms of vascular-related diseases often appear in the late stage and is associated with poor prognosis. Therefore the accurate risk stratification of patients with established coronary artery disease is essential as it will enable a personalised therapy of high risk individuals with novel therapies targeting disease progression. Evaluation of the lumen and vessel wall dimensions and plaque burden is essential for treatment planning and stratifying risk in patients with established coronary artery disease. This can be achieved using non-invasive techniques like computer tomography (CT), and invasive such as optical coherence tomography (OCT), coronary angiography (CA), and intravascular ultrasound (IVUS) (Rosales et al., 2009). CT has high sensitivity in the detection of coronary atheromatous lesions, however, it falls short in assessing the components of the plaque and does not provide accurate quantification of the plaque burden. CA provides a luminal projection

* Corresponding author.
E-mail address: qianni.zhang@qmul.ac.uk (Q. Zhang).
[2] These authors have contributed equally to this work.
[1] QCU-CMS; Leiden, University Medical Center, Leiden, The Netherlands.

https://doi.org/10.1016/j.media.2023.102922
Received 15 July 2022; Received in revised form 6 July 2023; Accepted 1 August 2023
Available online 5 August 2023

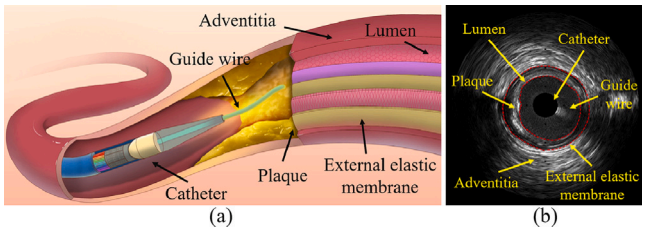


Fig. 1. (a) Demonstration of a IVUS catheter taking image on a vessel; (b) A typical IVUS image.

of the long axis of the vessel, which is the key information for the determination of the length and degree of stenosis of the coronary artery lesions. Nevertheless, it also cannot be used to assess plaque components and has limited accuracy. IVUS can obtain cross-sectional images of blood vessels, show the histomorphological characteristics of the vascular lumen, wall, and atherosclerotic plaque (Fig. 1), making it easier to guide coronary intervention with more accurate sizing of the length of lesion and dimensions of the vessel and guiding the need for plaque modification such as in the case of calcific lesions (Blanco et al., 2022). Thus, IVUS is currently one of the most effective imaging modalities that aids interventional cardiologists to diagnose and treat coronary disease. A visual illustration of IVUS capturing pullback and a resulting IVUS image are shown in Fig. 1.

Moreover, the segmentation of the IVUS images and the detection of the lumen and external elastic membrane and the quantification of the plaque burden also provides useful prognostic information and allows evaluation of the effects of novel pharmacotherapies on plaque evolution (Tufaro et al., 2023). Currently, the segmentation of IVUS is performed manually. An IVUS pull-back sequence contains a large volume of images, making the segmentation process extremely time-consuming. An automated solution for IVUS image segmentation is highly desirable.

Automated IVUS segmentation has been studied for decades and most of the published IVUS segmentation methods rely on hand-designed features (Balocco et al., 2003; O'malley et al., 2007; Ünal et al., 2008; Taki et al., 2008). In 2011, a dedicated IVUS segmentation dataset (IVUS-2011) was published in the "Lumen + External Elastic Laminae Border Detection in IVUS Challenge" in MICCAI (Essa et al., 2012), to offer a platform for validating and comparing systems developed for this task. This consists of around 500 groups of 5 consecutive frames, with about 25% of the data manually annotated. In recent years, deep learning technology has shown great success in the field of medical image segmentation tasks (Li et al., 2020; Sun et al., 2021; Cui et al., 2021). Several works have attempted to segment vessel walls of the IVUS sequences based on fully convolution networks (FCNs) (Blanco et al., 2022; Yang et al., 2018; Vercio et al., 2019; Xia et al., 2020), and have shown promising results. To our knowledge, most of the data-driven FCN based IVUS segmentation methods rely on a large number of training samples. Meanwhile, IVUS-2011 is the only published dataset to date, and its size is relatively small.

Several challenges remain in the IVUS segmentation task including the lack of well-annotated data, and the intrinsic variations, artifacts and large shadowed regions in vessels (Sheet et al., 2014). The shapes of vessel walls and plaques are various and complicated to outline. Shadowed regions are widely distributed throughout the blood vessels, along with artifacts, side-branches, and guide wire effect, as shown in Fig. 2(a)–(d), respectively. These intrinsic features make it extremely challenging even for human experts to precisely determine the EEM and lumen borders. The dark area is usually caused by the presence of calcified plaques, as shown in Fig. 2(a). The appearance of calcification in blood vessels leads to signal deflection and attenuation resulting in acoustic shadowing and the inability to visualise the area behind. Depicting the boundaries across the shadowed region requires referring to the other parts of the boundaries and making sensible inferences. Side-branch is another challenge for IVUS segmentation, as shown in Fig. 2(c). When a side-branch flows into the main branch at the junction, the two vessels slowly merge into one, the probe will detect both vessels without apparent boundary features. In this case, it is challenging to segment the EEM, because the identification of the vessel wall of the side-branches cannot rely on texture features, and thus requires leveraging the shape of vessels to infer the approximate position of the boundary. The blood vessels move back and forth due to the heartbeats, and the frame containing side-branches will appear intermittently for an extended period.

In order to tackle these challenges, we start by observing how cardiologists predict the boundaries in large shadowed regions where little visual information can be found. Human vision has excellent ability to perceive the visual organisations as a whole by mentally filling up the missing parts. This process is also guided by human's experience, which tends to interpret the incomplete structures in the way that it makes the most semantic sense. Thus, when a cardiologist tries to draw out the EEM and lumen boundaries, based on the boundary sections that are clear to depict, they can naturally complete the gaps in the boundaries, where a shadow is crossed. We mimic this process and propose a Perceptual Organisation-aware Selective Transformer framework for IVUS segmentation, namely, the POST-IVUS framework, which balances the requirement for highly accurate boundary detection based on suitable, representative dynamic visual features, and the need for logical predicting missing boundaries by simulating the perceptual organisation principle of human vision. Although human vision is imperfect and visual illusions commonly exist, particularly in shadowed regions, our POST-IVUS framework aims to closely resemble human predictions to provide the most reasonable estimations. By annotating a large number of images containing calcifications and improving our model's predictive capabilities, we acknowledge the potential inaccuracies in ground truth annotations while striving to advance segmentation performance in these challenging areas. The framework entails three main components:

The **temporal context-based feature encoders:** Two temporal context encoding schemes are designed to extract motion features that can effectively reveal the lumen border, including a rotational alignment encoder and a visual persistence encoder. The rotational alignment encoder eliminates the vessel's rotation motion introduced by heart beats and allows the encoder to focus on relevant vessel movement. The visual persistence encoder encodes the residual visual features in frame sequences that are particularly useful in identifying lumen borders in human visual examination. Our experiments show that the encoded features play an essential role in facilitating subsequent segmentation methods to accurately depict the boundaries.

A **Selective Transformer Recurrent U-Net with discriminator:** We propose Selective Transformer Recurrent U-Net (STR U-Net) as the backbone of POST-IVUS framework, based on our newly designed Selective Transformer Recurrent Residual (STRR) Block. The STRR Block is able to infer borders in dark regions while remain high accuracy in other areas. On top of that, we apply generative adversarial learning with a dedicated loss as the penalty standard, to effectively improve the ability of boundary inference. during the training stage, we added a discriminator to further strengthen the inference ability of STR U-Net.

A **Temporal constraining and Fusion module (TF Module):** The post-processing module specifically designed for IVUS segmentation task includes temporal constraint and late spatial augmentation fusion steps that enable improved annotation accuracy in side-branch and calcium areas where prediction is required and effectively eliminated incorrect annotation.

In this paper, experiments are performed and the results are reported to demonstrate the influence of the encoders, STR U-Net and TF modules. Overall, the POST-IVUS framework achieves excellent results on our private IVUS sequence dataset, and outperforms the state of the art on the public IVUS-2011 dataset. Especially, superior accuracy

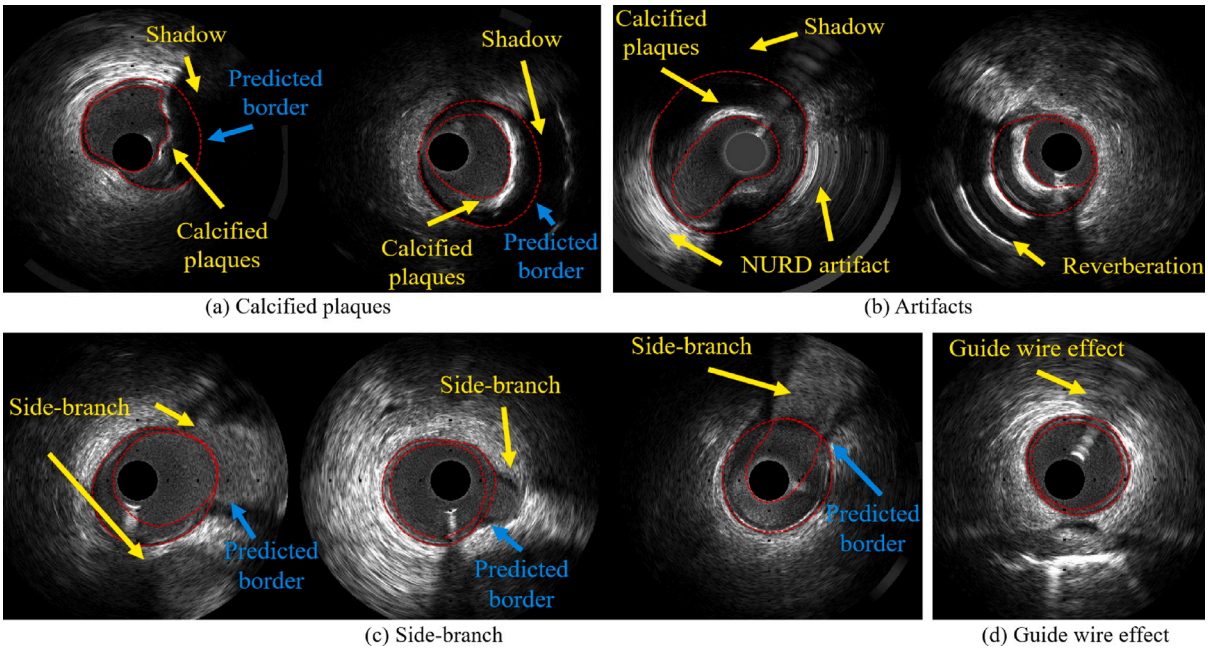


Fig. 2. Special cases of IVUS image sample, the red dotted line represents the expert annotation of EEM and lumen. Yellow arrow indicates the location of particular areas, and blue arrow indicates the location of predicted boundaries.

is achieved in the predicted EEM and lumen borders in the cases of side-branch, calcified plaques and artifacts. In lumen segmentation, our automatic method compares favourably to the inter-agreement among experts.

This method has been integrated into commercial Software QCU-CMS (Version 4.69, Leiden, University Medical Centre, Leiden, The Netherlands) along with an automated IVUS end-diastolic (ED) frame detection method. The average time for an experienced cardiologist to manually label EEM and lumen borders on all ED-frames of an IVUS pullback sequence, which normally consists of 5000–6000 frames, is 8-10 h. By using our proposed method, the same expert only needs about 10 min to briefly verify the automatically obtained segmentation boundaries, with few necessary corrections.

The main contributions of this research are summarised as follows:

- The POST-IVUS segmentation framework is proposed, which achieves over the state-of-the-art performance on IVUS-2011 dataset B, and exceeds the performance of human experts on our private NIRS-IVUS dataset.

- Two IVUS encoding schemes are proposed to extract the most relevant motion features by exploiting the temporal context information for lumen border prediction. IVUS motion features are enhanced by minimising the effect of irrelevant vessel rotation, and simulating visual residuals of human vision.

- A novel selective transformer (STRR) block is proposed. The resulting selective transformer recurrent residue U-Net (STR U-Net) encapsulates the excellent feature representations of recurrent residue blocks along with the enhanced prediction ability due to the large vision field introduced by Swin Transformer.

- An adversarial learning scheme is developed to guide and regulate the training segmentation models with perceptual organisation information. The adversarial objective is re-designed to force the model to simulate the human's visual perception ability of virtually completing semantic structures, and thus better predict boundaries in dark regions where little visual features can be captured.

- A new IVUS multi-class segmentation coding method and a loss that can utilise the topological relation between lumen and EEM borders, and a dedicated post-processing module, which can substantially eliminate network errors and bring more reasonable results in regions that require inference.

More background and related works are reviewed in Section 2. The proposed methods are described in detail in Section 3. The experiments are presented in Section 4, with a comprehensive evaluation in two datasets. An in depth discussion is provided in Section 5 and the paper concludes with Section 6.

2. Related work

Several types of traditional segmentation methods have been used to solve the IVUS vessel wall segmentation task, including knowledge-based methods (Sonka and Zhang, 1995), probabilistic and statistical methods (Gil et al., 2001), (Mendizabalruiz et al., 2010), filter based methods (Gil et al., 2006) and Wavelet-transform methods (Katouzian et al., 2010). Sonka and Zhang incorporated a priori knowledge of coronary artery anatomy and ultrasound image characteristics into the method for IVUS border detection (Sonka and Zhang, 1995). Gil et al. proposed a probabilistic approach to initialise a first approximation of an elliptical model which has a high probability of being close to the inner wall (Gil et al., 2001), then refine this ellipse using an adaptive threshold computed for each image. The same group later proposed blending advanced isotropic filtering operators and statistical classification techniques into a vessel border modelling strategy (Gil et al., 2006). Mendizabalruiz et al. introduced a sum of Gaussian functions that are deformed by the minimisation of a cost function formulated using a probabilistic approach to extract lumen contours (Mendizabalruiz et al., 2010). Katouzian et al. constructed the relative magnitude phase histogram of complex brushlet coefficients to determine luminal borders (Katouzian et al., 2010). All these above methods highly rely on feature engineering and lack of robustness and generalisability.

Active contour, also named snake or deformable contour, is wildly applied to solve IVUS segmentation tasks. Bourantas et al. used a deformable model and smoothed initial estimations for the external elastic membrane border (Bourantas et al., 2005). Iskurt et al. utilised an enhanced level set technique to derive the evolution of two coupled contours as the zero level sets of a single higher dimensional surface (Iskurt et al., 2006). Ginestar et al. used synthesised parametric curves within an image domain and allowed them to move towards the edges drawn by internal and external forces (Ginestar et al., 2014).

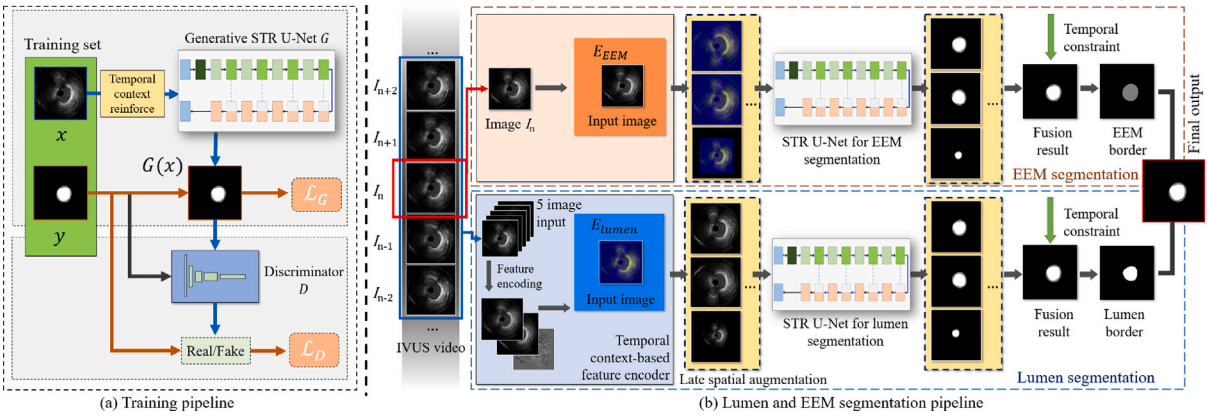


Fig. 3. The POST-IVUS framework architecture including a training pipeline (a) and a segmentation pipeline (b).

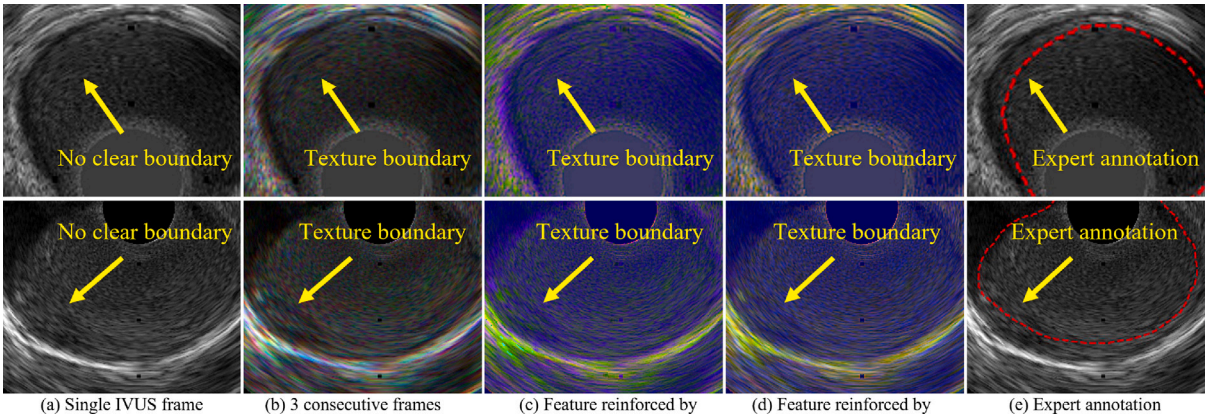


Fig. 4. Two example IVUS frames. (a) The lumen border cannot be seen in a single static frame. (b) When 3 consecutive frames are superimposed, the lumen border starts to appear. (c) The encoded feature map from the alignment encoder (AE), with a texture boundary of lumen shown. (d) The encoded feature map from the visual persistence encoder (VE), with a texture boundary of lumen shown. (e) The lumen border annotation by cardiology experts.

These active contour-based methods have good ability for denoising but are sensitive to the initial segmentation solutions.

In recent years, deep learning based segmentation methods surpass earlier conventional methods in segmentation accuracy (Destrempes et al., 2014). FCNs show excellent capabilities in tasks including lumen and out vessel wall segmentation in IVUS images. Destrempes et al. proposed a U-shape FCN architecture, called IVUS-Net (Yang et al., 2018), followed by a postprocessing contour extraction step, to automatically segment lumen and EEM boundaries of the human arteries. Li et al. developed an FCN model using three modified U-Nets to form cascaded networks to prevent errors in the detection of calcification on an IVUS dataset (Li et al., 2021). Bargsten et al. systematically investigated different capsule network architecture variants and improved the segmentation performances on IVUS sequences (Bargsten et al., 2021).

Generative Adversarial Networks (GAN) based methods attract a lot of attention in medical image segmentation (Goodfellow et al., 2014; Dai et al., 2017; Odena et al., 2017). Conditional GAN (cGAN) involves using prior information during the generation, driving cGAN to generate detailed segmentation results (Mirza and Osindero, 2014). The adversarial paradigm between the generator and discriminator enables boundaries to be predicted when visual features are lacking, and can accurately resemble the manually drawn boundaries. The GAN-based methods achieve remarkable results in various medical image processing tasks (Cui et al., 2021; Lei et al., 2020; Nie and Shen, 2020), but to our knowledge, its application to IVUS segmentation has not yet been considered.

Vision transformers show great potential in various vision tasks. By stacking multiple transformer blocks with vanilla attention, ViT processes non-overlapping image patches and obtains superior classification performance (Dosovitskiy et al., 2020). TransU-Net proposed by Jieneng Chen et al. in 2021 is the first network framework to apply Transformers to medical image segmentation (Chen et al., 2021). TransU-Net encodes the feature blocks output by the feature extraction network as the input sequence of Transformers to extract features taking into account the global context. Meanwhile, taking advantage of the structure of U-Net, the decoder upsamples the encoded features and then fuses them with high-resolution feature maps to achieve semantic segmentation of medical images. Compared to TransU-Net, which uses Transformers attention at the bottom of the network, UTnet (Gao et al., 2021) uses Transformers at the upsampling position. However, vanilla attention with quadratic complexity over the input length is hard to adapt to vision tasks with high-resolution images as input due to the expensive computational cost. To alleviate such issues, window-based attention is proposed to partition the images into local windows and conduct attention within each window to balance the performance, computation complexity, as well as memory footprint (Liu et al., 2021). This mechanism enables vision transformers to make great success in many downstream visual tasks. Recently, the attention mechanism has become a milestone technology for computer vision tasks. Some categories of attention modules achieve efficient performance for medical image segmentation tasks and indicate further directions for research on the IVUS segmentation task, such as SKNet (Li et al., 2019) and SENet (Hu et al., 2018). The attention-based paradigm is able to calibrate the weight for each class, but is not yet applied in the challenging task of IVUS segmentation.

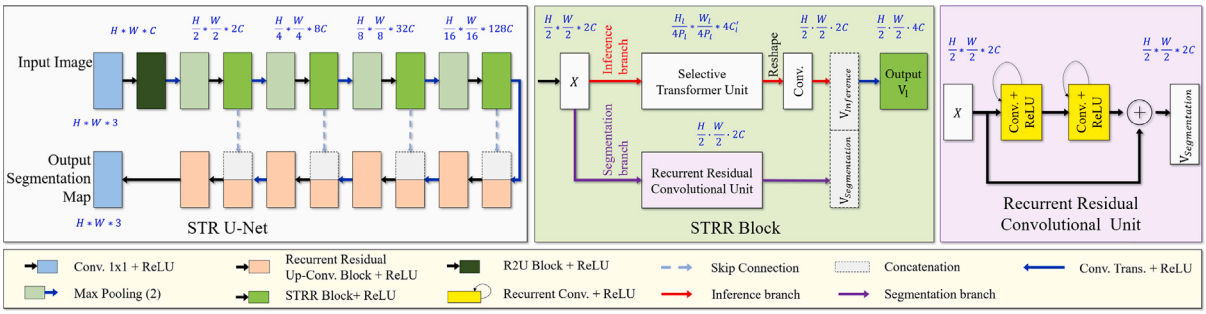


Fig. 5. The network structure of STR U-Net.

3. Methodology

IVUS segmentation is unconventional among medical image segmentation tasks. It shares the same requirement for accurate pixel-level segmentation masks. However, the common presence of dark regions where no image feature can be picked up, along with various types of artifacts, introduces unusual challenges, as illustrated in Fig. 2. While computer vision is being hampered by such challenges, these are not a problem to human experts, who have the superior ability in 'guessing' the boundaries across the shadowed regions, and not be confused by artifacts. This ability of human vision is summarised in Gestalt psychology, i.e., humans perceive the form of objects in a way that the overall organisation makes the most sense. More specifically, an important aspect of Gestalt visual perception is reification, in which the organisation as perceived contains more spatial information than what is actually present, and thus we find a near match and then mentally fill in the gaps in the actual structure. Moreover, humans can blend in their knowledge in this process, to help infer the perceptual organisation of the actual structure with missing parts. We observe how the expert cardiologists outline the lumen and EEM boundaries in IVUS frames where large shadows and artifacts are present, and design a series of methods that mimic their awareness of the visual perceptual organisation, and infer the complete, accurate boundaries. The aim is to perform a perceptual organisation-based inference, by leveraging visual information from neighbourhood regions, blending in prior anatomical knowledge, and conducting logical reasoning about the boundaries.

Inspired by the perceptual organisation awareness of human vision, we propose a perceptual organisation-aware selective transformer framework (POST-IVUS) for segmenting EEM and lumen boundaries in IVUS frames. In this section, we first give an overview of the framework architecture, and then present each main module in detail and discuss their design rationale.

3.1. POST-IVUS framework

The POST-IVUS framework consists of a training pipeline and a segmentation pipeline, as illustrated in Fig. 3. To tackle the lumen and EEM segmentation tasks effectively, the segmentation pipeline is the combination of two sub-pipelines, one dedicated for lumen border segmentation and the other for EEM. Since the detection of lumen border requires temporal features, we propose two temporal context-based feature encoders to exploit temporal information to strengthen the features for detecting the lumen border. We propose a new Selective Transformer Recurrent U-Net (STR U-Net) network as the segmentation backbone which imposes a Selective Transformer (ST) scheme on top of a recurrent residual convolution unit, in an overall U-shaped structure. Moreover, the adversarial training paradigm is utilised by adding a discriminator in the training stage to improve the perceptual inference ability in dark regions. Finally, temporal constraint and late spatial augmentation fusion are applied to improve the network robustness, and a temporal constraint is applied to eliminate erroneous areas.

3.2. Temporal context-based feature encoders

The lumen, in other words, the inner wall between the vessel and the blood, can hardly be seen on a single static IVUS frame, as illustrated in Fig. 4(a). To be able to identify the border, cardiologists tend to browse through a few frames before and after the target end-diastolic (ED) frame, and use the visual residue during the browsing to help visualise the lumen border. This is because the texture of vessels is relatively static, while the texture of the blood changes faster, so quickly browsing through a set of consecutive frames can help expose the lumen border. Similarly, in computer vision, the identification of lumen borders requires considering the visual cues in the temporal context, instead of looking at an ED frame statically. As shown in Fig. 4(b), when 3 consecutive frames are simply superimposed, a texture boundary representing the lumen border starts to show, in line with the expert annotation in Fig. 4(e). Here, we propose two temporal context-based feature encoders, to produce highly descriptive representations of the frames in lumen segmentation. The encoded features blend in the temporal context and enable capturing the subtle difference in texture changes of blood and vascular wall.

To take into account the temporal context, we consider two frames before an ED frame and two frames after as the keyframes. For an ED frame I_n, its group of key frames are denoted as $\{I_{n-2}, I_{n-1}, I_n, I_{n+1}, I_{n+2}\}$. Between any two frames I_q and I_p, a similarity score can be defined as follows:

$$\mathrm{Sim}\left(I_p, I_q, \theta\right) = \sum_{i}^{x}\sum_{j}^{y}\left[I_q(\mathrm{i},\mathrm{j}) - I_p^{\theta}(\mathrm{i},\mathrm{j})\right]^2, \tag{1}$$

where θ is a rotation factor that can be used to find the best rotational-aligned version of I_p to I_q.

To explore the motion feature for detecting lumen border, we proposed two temporal context-based feature encoders, namely, an alignment encoder (AE) and a visual persistence encoder (VE). The encoded features are used as the input to the STR U-Net.

Alignment Encoder: Each heartbeat introduces motion of the vessels, including a rotation and a back-and-forth motion. The rotation feature is not relevant to lumen border detection. Thus, to extract the actual relevant vessel movement features, the rotation change between every two consecutive frames need to be eliminated, i.e. the frames need to be aligned rotation-wise.

In AE, an original frame I_p is rotated between −20 and 20 degrees with a step size of 0.5 degrees around its centre and a rotated image is referred to as I_p^{θ}. Then, to calculate the best alignment $I_p^{\hat{\theta}}$ with respect to a reference frame I_q, the aim is to find a $\hat{\theta}$ such that:

$$\hat{\theta}_{(p,q)} = \arg\min_{\theta\in[-20°,20°]} \mathrm{Sim}\left(I_p, I_q, \theta\right). \tag{2}$$

The alignment of I_p to I_q is achieved by rotating I_p by $\hat{\theta}_p$ degrees, obtaining $I_p^{\hat{\theta}}$. For the sake of simplicity, we denote $I_p^{\hat{\theta}}$ as $\hat{I}_p$ in the following text. Based on $\hat{I}_p$, three feature channels are defined to represent the original image.

Denote Laplacian sharpening as $S(I) = \nabla^2 I(i,j)$, where $I(i,j)$ are a pixel's coordinates in image I. The first channel $AE_n^{(1)}$ is defined as:

$$AE_n^{(1)} = [\mathrm{S}(\hat{I}_{(n-2,n-1)}) + \mathrm{S}(\hat{I}_{(n-1,n)}) + \mathrm{S}\left(I_n\right) + \mathrm{S}(\hat{I}_{(n+1,n)}) + \mathrm{S}(\hat{I}_{(n+2,n+1)})]/5 \quad (3)$$

The second channel $AE_n^{(2)}$ is the original ED frame. The third channel $AE_n^{(3)}$ is defined follows. We derive the overall difference between the target frame I_n and the average of the three consecutive frames after Laplacian sharpening as D:

$$D = \frac{\mathrm{S}(\hat{I}_{(n-1,n)}) + \mathrm{S}\left(I_n\right) + \mathrm{S}(\hat{I}_{(n+1,n)})}{3} - I_n. \quad (4)$$

The values in D fall in the range of (−255,255), and are normalised into a range of (0,255). This normalised signal is the input to channel 3 $AE_n^{(3)}$. The three channels altogether form the output of the alignment encoder: $[AE_n^{(1)}, AE_n^{(2)}, AE_n^{(3)}]$.

Visual Persistence Encoder: This encoder's aim is to imitate the visual residuals of human vision for visualising the lumen border. Due to the visual residual effect, the relatively static blood vessel edges and the randomly changing blood texture can be separated by a visible border. We design two methods for visual persistence encoding, and corresponding outputs are fit into the first and third channel, while the middle channel is the original image.

The first channel in this encoder is designed to capture the temporal context in the neighbouring frames, by calculating the average of the five sharpened keyframes around an ED frame I_n. This channel is denoted as $VE_n^{(1)} = [S(I_{n-2}) + S(I_{n-1}) + S(I_n) + S(I_{n+1}) + S(I_{n+2})]/5$. The second channel simply takes in the original ED frame, $VE_n^{(2)} = I_n$.

The third channel $VE_n^{(3)}$ encodes the change data around I_n from its previous frame to its subsequent frame. We define $VE_n^{(3)'} = (I_{n-1} + I_n + I_{n+1})/3 - I_n$. $VE_n^{(3)'}$ is then normalised into the range of [0,255] to obtain $VE_n^{(3)}$, for capturing temporal changes by considering the previous, current, and subsequent frames, as well as filtering out high-frequency noise. This approach provides a more comprehensive understanding of the temporal dynamics and ensures that the output remains within the original image range, making it easier to interpret and utilise in the model.

The output of VE is $[VE_n^{(1)}, VE_n^{(2)}, VE_n^{(3)}]$.

3.3. STR U-Net: Selective transformer U-Net with recurrent residual blocks

IVUS segmentation requires not only highly precise, pixel-level delineation of lumen and EEM in conventional images, but also relies on further inferring the boundaries in challenging areas where visual cues are missing or misleading. For that purpose, the STR U-Net is proposed as the backbone of the POST-IVUS framework. This network inherits the strong segmentation power of U-Net, while it also possesses the inference ability that traditional segmentation networks do not have.

STR U-Net, akin to a U-Net, is a five-layer fully connected convolutional neural network. From the second layer onwards, a Selective Transformer Recurrent Residual (STRR) Block is incorporated as the encoder. The STRR Block integrates two feature extraction branches, one is called an 'inference branch' with a Swin Transformer, and the other is a 'segmentation branch' that entails a recurrent residual convolutional unit. The segmentation branch is designed to concentrate on pixel-level information, capturing the descriptive features in the local area for effective segmentation. However, relying on the local features is inadequate to solve the intrinsic problems in this task. As mentioned above, the IVUS images often include regions with obscured or absent visual information, such as the dark areas due to calcific plaques, or areas where side branches present, but the lumen and EEM boundaries run across such areas. We notice human vision system is particularly able to infer the missing part of boundaries in challenging areas by expanding the analysis to the whole images and mentally closing the gaps in the overall boundaries. The Swin Transformer is known to be powerful in capturing information from a larger receptive field. Therefore, we design the 'inference branch' with a Selective Transformer unit, to complement the segmentation branch and mimic the 'inference' process for the missing part of an organisational structure in human visual perception.

The two branches generate complimentary feature maps, which are merged in the succeeding layer. In the experiment evaluation, it is shown that this dual branch design in the STRR block is able to enhance the overall segmentation performance in various challenging scenarios in the IVUS segmentation task. A detailed illustration of the STR U-Net and the STRR Block structure can be seen in Fig. 5.

3.3.1. Segmentation branch

In the segmentation branch, shown as the purple flow in the STRR block in Fig. 5, we replace the convolution layer in the traditional U-Net with a recurrent residual convolutional unit (RRCU). Recurrent units improve the memory capacity of the network and help learn better feature representations for vision-related tasks. RRCU has been proven effective and efficient in training deep neural networks in Alom et al. (2018).

In the l_{th} layer of STR U-Net, at time step t, the output of the improved residual network is denoted as $O_{ijkl}(t)$, which corresponds to the pixel coordinates i,j in the k_{th} feature map. The mathematical expression of the output is:

$$O_{ijkl}(t) = (w_k^f)^T * x_l^{f(i,j)}(t) + (w_k^r)^T * x_l^{r(i,j)}(t-1) + b_k. \quad (5)$$

$x_l^{f(i,j)}(t)$ and $x_l^{r(i,j)}(t-1)$ are the inputs to the standard convolution layer and the recurrent convolutional layer, respectively. w_k^f represents the weights of the standard convolutional layer of the k_{th} feature map; w_k^r represents the weights of recurrent convolutional layer of the k_{th} feature map; and b_k is a bias. The final output of segmentation branch $V_{Segmentation}$ is represent as:

$$V_{Segmentation} = max(0, O_{ijkl}(t)) \quad (6)$$

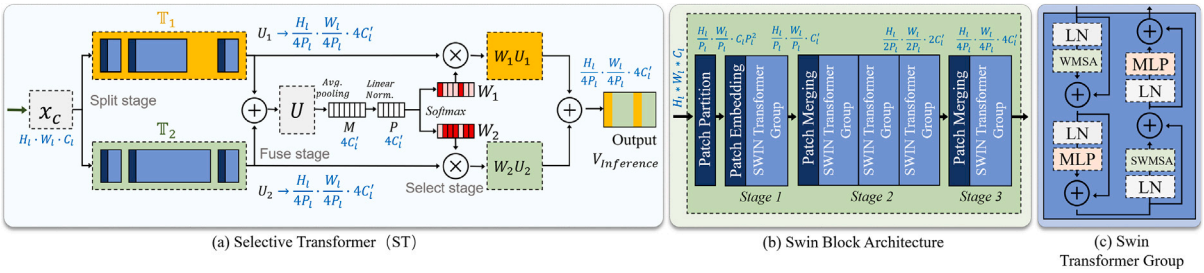
3.3.2. Inference branch

The inference branch is indicated in red in the STRR Block in Fig. 5. The field size is a crucial factor to the inference ability of a segmentation model. In order to infer the boundaries in regions where little information presents, i.e. the dark areas and side-branches areas, it is necessary to involve the visual information from adjacent context areas by mimicking the perceptual organisation property of human vision. We propose a Selective Transformer structure to perform segmentation inference by taking into account larger visual fields, inspired by the SKNet (Li et al., 2019). The structure of the selective transformer is depicted in Fig. 6(a).

The global self attention mechanism offers a large field of view. Therefore, we use Swin Transformer as the backbone of the inference branch (Liu et al., 2021). Swin Transformer introduces connections between areas by dividing images into different patches and calculates hierarchical representation. Then, key neighbour patches can be exploited when speculating patches have insufficient information. The selective transformers is designed to find patches containing relevant context information in an adaptive way.

Here, a Swin Transformer group is built to explore patch-based relationships on two split feature maps from different convolution kernels. There are three stages in the selective transformer: splitting stage, fusing stage and selecting stage.

Splitting stage: For a given feature map x_l of size $[H_l, W_l, C_l]$, by default we first place two Swin blocks $\mathbb{T}_1 : x_l \rightarrow U_1 \in \mathbb{R}^{H \times W \times C}$ and $\mathbb{T}_2 : x_l \rightarrow U_2 \in \mathbb{R}^{H \times W \times C}$. The structure of a Swin block is shown in Fig. 6(b). The input feature map in layer l of the STR U-Net are defined as $[B, C_l, H_l, W_l]$, where B denotes the batch size, C, H and W denote the feature channels, height and width of the input feature map. To allow a Swin block to serve its purpose, the input feature map is



(a) Selective Transformer (ST) (b) Swin Block Architecture (c) Swin Transformer Group

Fig. 6. Key modules in the STRR block. (a) The proposed selective transformer structure in STRR block's inference branch. (b) The structure of a Swin block, i.e., $\mathbb{T}_1$ or $\mathbb{T}_2$ in the splitting stage of ST. (c) The structure of each Swin Transformer group.

first partitioned into non-overlapping patches. We extract the high-level features as 'tokens' from the original patches.

Denote the side length of a square input image as L, the patch size is defined as:

$$P_l = \begin{cases} \frac{L}{2^{l} \times 10} & l = 2, 3, 4 \\ \frac{L}{2^{l-1} \times 15} & l = 5. \end{cases} \tag{7}$$

In stage 1 of the Swin block in Fig. 6(b), the number of tokens remains $(\frac{H_l}{P_l} \times \frac{W_l}{P_l})$. The attention heads are set to 4 for each layer. The patch embedding layer resizes the feature dimension to $C'_l = 3 \times 2^{8-l}$, while the feature map becomes deeper and smaller in each layer.

Then in stage 2, the patch merging layer concatenates features of neighbouring 2×2 patches in each group, so the output dimension becomes to $2C'_l$, and the number of tokens reduces to $\frac{H_l}{2P_l} \times \frac{W_l}{2P_l}$. Then, a few Swin Transformer groups are applied for feature transformation. Finally, the same process is repeated in stage 3, but with only one Swin Transformer group. The output resolution reduces to $\frac{H_l}{4P_l} \times \frac{W_l}{4P_l}$. Thus, the shape of the Swin block's output in layer l is $[\frac{H_l}{4P_l}, \frac{W_l}{4P_l}, 4C'_l]$. So the dimensions of the hidden states in each layer l is $4C'_l = 4 \times 3 \times 2^{8-l}$.

In each Swin Transformer group, as illustrated in Fig. 6(c), the Window-based Multi-head Self Attention module (WMSA) evenly divides an input feature map into windows to reduce computational complexity, on top of the original Multi-head Self Attention module (MSA) (Vaswani et al., 2017). The Shifted Window-based Multi-head Self Attention module (SWMSA) then changes the order of the embedded patches in WMSA. The process always starts by going through a LayerNorm (LN) layer before every MSA or a 2-layered Multi-Layer Perception (MLP) module. The first WMSA module partitions an original feature map of $[\frac{H_l}{4P_l}, \frac{W_l}{4P_l}, 4C'_l]$, and window size is set to 7 for $\mathbb{T}_1$ and 9 for $\mathbb{T}_2$. The second MSA module (SWMSA) shifts the output of the previous layer, rights-wards and down-wards for half a window size. The process in a Swin Transformer group can be formally defined as the following:

$$\begin{aligned} z_1 &= WMSA(LN(x_l)) + x_l, \\ z_2 &= MLP(LN(z_1)) + z_1, \\ z_3 &= SWMSA(LN(z_2)) + z_2, \\ z_4 &= MLP(LN(z_3)) + z_3, \end{aligned} \tag{8}$$

where x_l is the input to the Swin Transformer group, z_1, z_2, z_3 and z_4 is the output features of the WMSA module, the MLP module, the SWMSA module and the final output of Swin Transformer group, respectively. In the Swin block of $\mathbb{T}_1$, there are two Swin Transformer groups in its stage 2 and that of $\mathbb{T}_2$ contains three. For thicker vessels, a big window size in $\mathbb{T}_1$ can cover a larger area, adequately exploring the characteristics of both EEM and lumen boundaries. Meanwhile, more fine-grained information is required for segmentation in small vessels, assisted by inference in the areas with guide wire artifacts and small plaques. Therefore, $\mathbb{T}_2$ with a smaller window size can focus on collecting useful features for the inference process.

Fusing stage: A series of gates are employed to control the weights of multiple feature maps in order to find the most relevant patches for inference. We first calculate a fused feature map U from the two branches: $U = U_1 + U_2$.

Then the global information is embedded by global average pooling to generate the statistics feature map matrix channel-wise as M. For every element in feature channel k, we compress the feature map U_k to M_k by down-sizing U_k through spatial dimensions $\frac{H_l}{4P_l} \times \frac{W_l}{4P_l}$:

$$M_k = \frac{1}{\frac{H_l}{4P_l} \times \frac{W_l}{4P_l}} \sum_{i=1}^{\frac{H_l}{4P_l}} \sum_{j=1}^{\frac{W_l}{4P_l}} U_k(i, j) \tag{9}$$

The size of M is $[1 \times 4C'_l]$ in layer l. Then a linear layer and a normalisation layer are applied on M, to generate a feature map P of shape $[1 \times 4C'_l]$.

Selecting stage: Here, a softmax operation is applied to feature map P of each channel. This step enables the network to select the most relevant patches for inference. The kth feature in P are denote as P_k. For each feature map U_1 and U_2, the weight W of feature P is calculate as:

$$W_{1,k} = \frac{e^{P_{1,k}}}{e^{P_{1,k}} + e^{P_{2,k}}}, W_{2,k} = \frac{e^{P_{2,k}}}{e^{P_{1,k}} + e^{P_{2,k}}}, \tag{10}$$

where $W_{1,k} + W_{2,k} = 1$. Then we can obtain the overall feature map V_k:

$$V_k = W_{1,k} \cdot U_{1,k} + W_{2,k} \cdot U_{2,k}, \tag{11}$$

where $V_k \in \mathbb{R}^{H \times W}$, and $V = [V_1, V_2, \ldots, V_k]$. Each output feature map V_k is reshaped to $[\frac{H}{2}, \frac{W}{2}, 4C'_l]$. At last, the features are reduced by a convolutional layer to obtain the final output of the inference branch $V_{Inference}$, of shape $[\frac{H}{2}, \frac{W}{2}, 2C]$.

The overall feature map of the STRR block V_l for layer l is generated by concentrating $V_{Inference}$ and $V_{Segmentation}$, as illustrated in Fig. 5.

3.4. Adversarial learning

To further improve the capability of perceptual organisation based inference in segmentation, we employ the generative adversarial learning scheme to guide and regulate the training of segmentation models. The learning process will force the model to simulate the way cardiologists predict sensible boundaries in shadowed, confusing regions and outline the complete inner and outer vessel walls. When the network generates unreasonable segmentation, the discriminator will be able to identify it and penalise the aberrant area with its loss. As shown in Fig. 7, a second area is mis-detected due to the influence of pericardium and side-branch, and this false area will be suppressed by adversarial learning.

An adversarial learning based segmentation network includes two main parts, as depicted in Fig. 3(a): a generative network G and a discriminative network D. G is used to generate segmentation masks, and D is used to judge whether a mask is real or generated by the generative network. Specifically, given an image I, the generative network aims to learn a mapping from I to a mask M, namely, $G : I \rightarrow M$. In the setting of adversarial learning, the generator G aims to

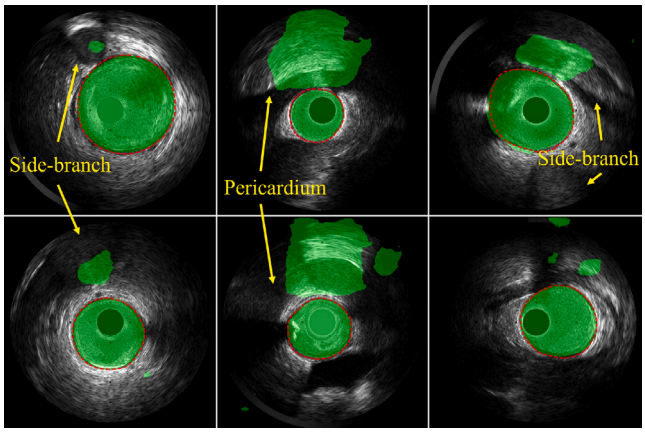


Fig. 7. Some examples of U-Net predictions for EEM regions. Green area: predicted results, red dash lines: expert annotations. The main reason for detecting multiple regions is due to the presence of side-branch and pericardium.

generate accurate segmentation results to fool the discriminator D so that it cannot distinguish a generated mask M from the corresponding real mask $\hat{M}$. The objective can be expressed as:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}\left[\log(D(M)\right] + \mathbb{E}\left[\log\left(1 - D(G(I)\right)\right] \tag{12}$$

$\mathbb{E}[\cdot]$ represents the expected value over all data instances. We also apply $\mathcal{L}_{L1}$ (L1 loss) to train the generative network G:

$$\mathcal{L}_{L1}(M, \hat{M}) = \|M - \hat{M}\|_1 = \|G(I) - \hat{M}\|_1 \tag{13}$$

Given a coefficient λ between $\mathcal{L}_{L1}$ and $\mathcal{L}_{GAN}$, the final objective is:

$$\mathcal{L} = \min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L1}(G(I), \hat{M}). \tag{14}$$

By adding the discriminator to the POST module, the network will force STR U-Net to resemble the most likely annotation of human experts and generate reasonable edges in areas where inference is needed. At the same time, it can strengthen the correlation between EEM and lumen boundary and ensures the inference is reasonable. In the POST-IVUS framework, the discriminator is essential for enhancing segmentation accuracy by guiding the generative network to produce more accurate and realistic segmentation masks. We use the discriminator during training in all instances, except for select ablation study experiments. In these ablation studies, the discriminator is excluded to assess individual components or variations within the framework, allowing us to understand their contributions to overall segmentation performance. Through the consistent application of the discriminator during training, the generative network is encouraged to generate segmentation masks that closely mimic expert annotations, leading to improved segmentation outcomes.

3.5. Spatial constraint and losses

Lumen and EEM represent the inner and outer walls of the vessel, respectively. Therefore, the lumen region is always enclosed in the EEM region. This prior knowledge can be exploited in network design as a spatial constraint for defining both boundaries. Thus, we design a training strategy in which both segmentation ground truth are considered jointly, i.e., for lumen segmentation, both lumen and EEM masks are used for training, and vice versa. In the learning process, the two predicted segmentation boundaries need to satisfy the constraint that lumen is enclosed inside the EEM border.

For multi-class segmentation tasks, two types of losses are commonly used: distribution-based losses, such as $\mathcal{L}_{CE}$, or region-based losses, such as $\mathcal{L}_{Dice}$ and $\mathcal{L}_{IOU}$. The annotation of each class occupies a channel, and a total of $n + 1$ channels are required for n classes. In this way, each channel is independent, and the final output is obtained by the softmax of all channels. This approach does not impose the relationship between EEM and lumen borders as the mutual influence between channels is small, and thus the resulting boundaries unavoidably contain errors. To address this issue, we put these two categories into one channel for training, where the vascular position is grey and the lumen interior is white. $\mathcal{L}_{L1}$ is utilised, as it does not require softmax operation. In this way, the $\mathcal{L}_{L1}$ features can generate sharp boundaries (Isola et al., 2017), the segmentation accuracy of lumen borders is improved. The experimental results also prove that IVUS segmentation can benefit from the spatial relationship between the two classes. In dark regions, as the shape of vessels defined by the EEM is relatively fixed, the lumen boundary can be better demarcated by taking advantage of the spatial relation constraint.

3.6. Temporal constraint and fusion module

A temporal constraining and fusion (TF) module is designed as a post-processing stage in the POST-IVUS framework. Due to the presence of side-branches, noise and artifacts in IVUS videos, the segmentation results are often unstable, and the predicted region masks may differ greatly from the target segmentation regions. To solve this series of problems and improve the performance robustness, we propose a temporal constraining and fusion module, which includes two parts: temporal constraint and late spatial augmentation fusion.

3.6.1. Temporal constraint

First, we proposed an IVUS temporal context constraint approach to eliminate false positives caused by side-branch and noise. The proposed adversarial learning approach for segmentation is capable of eliminating the false positive regions detected due to the presence of side-branch and pericardium, as shown in Fig. 7. However, in some cases, small regions of a few pixels, as an effect of noise, will remain. To address this issue, we propose to seek for additional temporal context information from neighbouring frames to help eliminate the unwanted areas in the mask. This is designed as a post-processing step, in which the temporal context constraint is applied on the initial masks obtained by the segmentation networks, identifying and keeping the most likely correct regions according to the temporal context, while eliminating the rest that are likely to be false positives due to side-branches and noise.

More specifically, if the segmentation mask M_n of frame I_n contains more than one disconnected region $R_t, t > 1$, define a function $\text{Area}\left(R_t\right)$ that calculates the area of a region, then the final mask M'_n can be express as:

$$M'_n = \begin{cases} \arg\max_{R_t \in M_n} \left\{\text{Area}\left(R_t\right)\right\}, n = 1 \\ \arg\max_{R_t \in M_n} \left\{\text{Area}\left(R_t \cap M'_{n-1}\right)\right\}, n > 1 \end{cases} \tag{15}$$

Now, the final mask M'_n of frame I_n contains only one EEM region.

3.6.2. Spatial augmentation fusion

To further enhance the segmentation accuracy, the robustness against morphological variances and stability in the challenging areas, we developed a spatial augmentation fusion step in post-processing. Based on a segmentation mask M'_n for ED frame I_n, the centre of the vessel, i.e., the masked EEM region, can be calculated. In the first type of augmentation, the ED frame I_n is rotated by a certain degree with respect to its image centre to generate a few new versions of the frame. In the second type of augmentation, the frame is shifted so that the vessel centre is aligned with the image centre. Then the shifted frame is rotated by a certain degree with respect to its image centre to generate a few new versions of the frame. In the third type of augmentation, the original and shifted frames are transposed. In total ten varied frames are obtained including the original.

First, a frame I_n is translated by aligning the vessel centre to the image centre, and the resulting image is denoted I'_n.

Then, 10 varied versions of the original image I_n is acquired by the following operations: $\{I_n, I'_n, I_n^{90°}, I_n'^{90°}, I_n^{180°}, I_n'^{180°}, I_n^{270°}, I_n'^{270°}, [I_n]^T, [I'_n]^T\}$, where T represents the transpose operation and I_n^{θ} represents I_n rotated by θ degrees. The missing corners generated by frame translation and rotation are all zero-padded so that the images keep the same scale.

For each of the images, a segmentation mask can be acquired using the above described method. The resulting 10 masks are reversely rotated, translated and transposed to obtain masks at the same position as that of the original image.

Define M_n^{sum} as the overlay of the 10 masks in the form of a 480×480 matrix, in which every pixel is a vote value telling how many masks include this pixel inside the segmentation region. Apply threshold H on the vote values and a final segmentation result $\hat{M}_n$ is achieved for image I_n. For every pixel $M_n^{\text{sum}}(x, y)$ in M_n^{sum}:

$$\hat{\mathrm{M}}_n(x, y) = \begin{cases} 1, & M_n^{\text{sum}}(x, y) \geq H \\ 0, & M_n^{\text{sum}}(x, y) < H \end{cases} . \tag{16}$$

In the experiments of this paper, empirically, the thresholds are set to $H = 7$ for EEM and $H = 5$ for lumen.

4. Experiments and results

In this section, we present the experiment results with quantitative and qualitative evaluations. The influences of different method selections and setups are analysed.

4.1. Evaluation datasets

The POST-IVUS framework is evaluated on a private dataset (NIRS-IVUS) and a public dataset (IVUS-2011). These datasets are described in the following.

4.1.1. NIRS-IVUS dataset

We analysed IVUS data from a total of 70 patients who participated in the "Evaluation of the efficacy of computed tomographic coronary angiography in assessing coronary artery morphology and physiology" study (NCT03556644). In the major epicardial vessels in these 70 patients, 197 near-infrared spectroscopy (NIRS)-IVUS sequences were captured by Infraredx 2.4F 50 MHz Dualpro system. The pullback was performed by an automated device with a fixed speed of 0.5 mm/s, and the frame rate of the IVUS video was 30 frames per second. The frame resolution is 480×480 pixels.

In total 197 vessels were assessed providing a total of 26,678 end diastolic frames, and annotated by four experts over a period of two years. According to our dataset acquisition standards and previously published guidelines (Mintz et al., 2001), we excluded images in which the EEM borders in obscured regions for an arc $> 90°$. This dataset, reported in the literature (Bajaj et al., 2021), is currently the largest dataset for training lumen and EEM border segmentation. The amount of data and the associated data acquisition workload far exceeds the magnitude of those in existing datasets based on public information. The statistics and division of the NIRS-IVUS test set are shown in Table 1.

In this dataset, we use 23,774 fully annotated ED frames from 177 vessels for training. The neighbouring frames before and after the ED frames are needed in the lumen boundary segmentation, therefore, a total of 118,870 frames are involved in training. For independent testing, 2,437 tagged ED frames from 20 vessels are used, including 241 side-branch and 229 calcification cases. The training and test sets are collected from different patients. On the test set, two experienced cardiologists performed annotation to evaluate the inter-observer variability. Finally, a senior expert with more than 10,000 h of experience in this work selected the best label and manually modified it to obtain the gold standard.

Table 1
Statistics and division of the NIRS-IVUS test set.

	Normal	Side-branch	Calcium
Test set	1987 (81.5%)	241 (9.9%)	229 (9.4%)

4.1.2. IVUS-2011 dataset

This dataset was first published with the IVUS 2011 open challenge and is widely used in the literature (Balocco et al., 2014). It consists of two subsets, the first having 77 in-vivo coronary artery frames captured using a 40 MHz Atlantis SR40 Pro catheter from 22 patients, while the second has 435 in-vivo coronary artery frames are taken from 10 patients by a Si5 Volcano Corporation device equipped with a 20 MHz Eagle Eye probe. The latter is one of the most widely used in the IVUS segmentation field. Following most of the research, this paper only considers the second subset for experiments and evaluation. Among the 435 frames in the dataset, each sized 384×384 pixels, 60 include the presence of bifurcation (side-branches), 94 have side vessels and 108 contain shadow artifacts (mainly due to calcific plaques). According to the challenge requirements, 109 frames are used as the training set, and the remaining 326 are used as the testing set. The ground truth was established by four clinical experts, and this dataset also provides four adjacent frames for each end-diastolic frame.

4.2. Evaluation metrics

For the two datasets, different evaluation criteria are used. For IVUS-2011 dataset, there are three widely used measurements: Hausdorff distance (HD), Jaccard measure (JM) and percentage of area distance (PAD). To analyse the results in further detail, four parameters are used for the NIRS-IVUS dataset: Hausdorff distance, mean distance (MD), Dice index (Dice) and Jaccard measure. Wherein, HD and MD are distance-based measurements; JM, PAD and Dice are area-based measurements. For the NIRS-IVUS dataset, the actual distance for one pixel is 0.0222 mm (22.2 μm), and 0.026 mm (26 μm) for IVUS-2011 dataset B. The distance in NIRS-IVUS dataset is measured in micrometres (μm), and millimetres (mm) for IVUS-2011 dataset. The Dice, PAD and JM are measured in percentage.

Hausdorff distance estimates the maximum distance between a point in the ground truth boundary and its nearest point in the predicted boundary. Denote the predicted boundary as B_{auto} and the manually annotated boundary B_{man}. The Hausdorff distance HD is calculated as:

$$HD = \max_{p \in B_{\text{auto}}} \left\{ \min_{q \in B_{\text{man}}} \{\text{dist}(p, q)\} \right\}, \tag{17}$$

where p and q are a point in B_{auto} and in B_{man}, respectively, and $\text{dist}(p, q)$ is a function that calculates the distance between the two points.

Similarly, the mean distance is often used to measure the accuracy of a segmentation boundary, calculated as:

$$MD = \left(\sum_{p \in B_{\text{man}}} \min_{q \in B_{\text{auto}}} \{\text{dist}(p, q)\} \right) / \left| B_{\text{man}} \right|, \tag{18}$$

where $\left| B_{\text{man}} \right|$ is the number of pixels in manually annotated boundary.

Denote the area of a manually annotated segmentation as A_{man}, and that of the model-predicted segmentation as A_{auto}. A Jaccard measure is also referred to as the Intersection over Union (IoU), and it is calculated as:

$$JM = \frac{\left| A_{\text{man}} \cap A_{\text{auto}} \right|}{\left| A_{\text{man}} \cup A_{\text{auto}} \right|} \times 100\% \tag{19}$$

The Dice coefficient, also known as the dice similarity coefficient or F1 score, is denoted as:

$$Dice = \frac{2 \left| A_{\text{man}} \cap A_{\text{auto}} \right|}{\left| A_{\text{man}} \right| + \left| A_{\text{auto}} \right|} \times 100\% \tag{20}$$

Table 2
EEM segmentation results on the NIRS-IVUS dataset, with the comparison of losses. All results are generated based on U-Net.

	All				Normal				Side-branch				Calcium			
	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM
$\mathcal{L}_{Boundary}$	184.6	64.4	97.532	95.411	139.0	50.3	97.964	96.163	515.3	185.7	94.335	89.983	257.4	63.9	96.654	93.754
$\mathcal{L}_{Boundary} + \mathcal{L}_{Dice}$	183.3	53.6	97.422	95.210	138.7	42.8	97.831	95.932	491.1	136.5	94.409	90.030	274.5	67.0	96.585	93.599
$\mathcal{L}_{IOU}$	173.2	49.2	97.802	95.801	119.8	36.9	98.246	96.585	531.4	135.2	**94.895**	90.715	293.8	72.7	96.666	93.735
$\mathcal{L}_{CE}$	172.8	46.1	97.868	95.927	119.3	33.5	98.298	96.684	539.9	134.1	94.886	**90.743**	278.8	69.4	**96.944**	**94.212**
$\mathcal{L}_{Tversky}$	174.2	46.7	**97.896**	**95.992**	121.2	34.2	**98.360**	**96.809**	536.5	132.8	94.801	90.607	281.4	70.4	96.747	93.904
$\mathcal{L}_{Dice}$	168.2	46.0	97.811	95.810	**115.9**	34.4	98.252	96.589	523.4	125.2	94.855	90.652	281.5	69.5	96.710	93.796
$\mathcal{L}_{L1}$	**164.3**	**42.3**	97.369	95.096	124.2	**33.2**	97.743	95.767	**435.9**	**102.0**	94.690	90.408	**251.7**	**62.8**	96.584	93.563

Table 3
EEM segmentation by using only lumen annotations and lumen annotations in training. All results were generated based on U-Net + Dis. + TF ($\lambda = 100$), $\mathcal{L}_{L1}$, NIRS-IVUS dataset.

EEM segmentation	**Input**	**HD**	**MD**	**Dice**	**JM**
1 Frame	Lumen	166.4	43.7	97.686	95.563
1 Frame	Full	**143.8**	**37.6**	**97.992**	**96.123**
Lumen segmentation	**Input**	**HD**	**MD**	**Dice**	**JM**
Aliment Encoder	Lumen	238.6	73.4	94.891	90.470
Visual Persistence Encoder	Lumen	220.0	67.8	95.230	91.066
Aliment Encoder	Full	199.1	60.4	95.761	95.761
Visual Persistence Encoder	Full	**194.9**	**57.6**	**95.897**	**95.897**

The percentage of area distance is defined as:

$$PAD = \frac{|A_{\text{man}} - A_{\text{auto}}|}{A_{\text{man}}} \times 100\% \tag{21}$$

In order to estimate the level of agreement between different observers, and between the model and observers, we employ the modified Williams Index (mWI) and 95% confidence interval for the modified Williams index (Chalana and Kim, 1997). WI is a metric designed to measure the ratio of the observer-to-observer or computer-to-observer agreement. For two annotations by the i_{th} and j_{th} observers, $i, j \in [0, n]$, the proportion of disagreement is denoted as $D_{i,j}$. For HD, MD and area, the proportion of disagreement $D_{i,j}$ equals to the difference between the distance measurements of observers i and j. For the Dice index and Jaccard measurement, the proportion of agreement between observers i and j are denoted as $P_{i,j}$, and $D_{i,j} = \frac{1}{P_{i,j}}$. The modified Williams index is defined as:

$$mWI' = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{D_{0,i}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{i':j\neq i}\frac{1}{D_{i,j}}}, \tag{22}$$

for a total number $(n + 1)$ observers, from 0 to n. The mWI compares the degree of agreement between the computer-generated boundaries and each observer's manual annotations against the level of agreement among the observers themselves. When the mWI is greater than 1, it implies that the computer-generated boundaries are more agreeable to each expert compared to their agreements, thus signifying that the computer-generated results better represent the consensus among the experts.

4.3. Experimental setting

All the experiments are carried out in the same hardware and software environment: eight Lenovo SR670 servers, each include 2 of 24 cores Intel Xeon Platinum 8268 CPU, 384 GB memory and four of NVIDIA A100 40 GB GPU. We build the entire project based on Python 3.8 and PyTorch 1.8.1.

The processing speed after loading the pullback video is 2 frames per second. The optimiser is Adam (Kingma and Ba, 2014) and the learning rate is 0.00001. The model's weight is initialised randomly, and the training epoch has been set to 200. For the two datasets, the input sizes have been configured to 480 ×480 and 384 × 384, respectively. A full segmentation of the test set is performed at every epoch, and the batch size is set to 32. The training data is enhanced by flipping, rotating, translating, panning, zooming, adding Gaussian noise, elastic distortion, and randomly adjusting image brightness.

4.4. IVUS segmentation results

First, we explore the performance differences of annotation methods, losses and models in the NIRS-IVUS EEM border segmentation task, to find the best parameter combination. We also compare our proposed STR U-Net with other well-known segmentation networks. Then NIRS-IVUS lumen segmentation results are reviewed to explore the best encoder for lumen segmentation. We then compare the best method with the annotations of two experts to evaluate Computer-to-observer and inter-observer variability. Finally, we compare our proposed method with other publicly available methods since 2018 on the IVUS-2011 dataset B.

4.4.1. Loss and spatial constraint

Using both lumen and EEM annotations in the EEM segmentation pipeline can effectively improve segmentation performance, and vice versa. Based on Table 3, in each segmentation pipeline, the performance is improved due to the existence of both types of annotations. This is especially evident with the segmentation on the lumen border, as the EEM boundary can effectively constrain the lumen boundaries on side-branch areas. At the same time, We quantitatively evaluate the performance of all segmentation losses on the EEM segmentation task in Table 2 to determine the extent to which each loss can take advantage of this constraint. Fig. 9 presents sample masks for qualitative assessment. In terms of critical indicators such as Hausdorff distance and mean distance, $\mathcal{L}_{L1}$ outperforms other losses significantly. It produces more stable borders and fewer errors. In the side-branch and calcium cases that require prediction, $\mathcal{L}_{L1}$ can generate more reasonable and stable predicted boundary that resembles the ground truth. As described in Section 3.5, the use of $\mathcal{L}_{L1}$ allows the training of both classes to be carried out in a single channel simultaneously, the inherent spatial constraints between EEM and lumen boundaries can be adequately taken into account. This is the main reason why $\mathcal{L}_{L1}$ loss leads to superior results compared to other losses. Both the EEM and lumen annotations are jointly applied in training. In particular, the stable EEM border can constrain the position of lumen boundaries, and enhance the performance of lumen segmentation.

For the other losses, $\mathcal{L}_{Boundary}$ (Kervadec et al., 2019), $\mathcal{L}_{CE}$, $\mathcal{L}_{IOU}$, $\mathcal{L}_{Tversky}$, and $\mathcal{L}_{Dice}$, three-channel masks are required. As illustrated in the examples in Fig. 9, the resulting boundaries are less satisfactory and are easier to be influenced by shadows and side-branches.

Overall, this set of experiments verifies that applying the spatial constraint and placing all boundaries in one channel leads to more reasonable results. Therefore, $\mathcal{L}_{L1}$ is selected as the main loss in the POST-IVUS framework, and is applied in all subsequent experiments.

4.4.2. NIRS-IVUS EEM segmentation

In this study, we design the STR U-Net, which combines the strengths of the selective transformer (ST) and the Recurrent Residual Convolutional Unit (RRCU). While both ST and RRCU are capable of extracting hierarchical information, they serve complementary roles in the segmentation process. On the one hand, ST captures long-range dependencies and global context, which is particularly beneficial for handling complex structures and larger visual fields. On the other hand, RRCU focuses on learning local features and maintaining spatial information, which is crucial for detailed boundary delineation.

Table 4
EEM segmentation results on NIRS-IVUS, comparison between the proposed ST U-Net, STR U-Net and previous segmentation models.

Network	All				Normal				Side-branch				Calcium			
	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM
SegNet	182.7	76.6	97.629	95.544	125.5	50.2	98.099	96.354	496.8	188.8	94.704	90.515	451.7	258.0	95.596	92.183
PSPNet	138.0	39.1	97.972	96.098	102.9	30.7	98.311	96.699	383.8	98.4	95.590	91.922	218.3	59.4	97.094	94.521
DeepLab V3+	143.0	39.4	97.970	96.076	105.5	30.9	98.322	96.712	393.5	94.7	95.677	91.961	236.0	63.6	96.957	94.236
ENet	151.8	41.3	97.698	95.610	114.0	33.1	98.054	96.242	414.8	98.0	95.143	91.143	240.8	64.0	96.825	94.010
GCN	141.2	39.0	97.909	95.979	106.8	31.3	98.278	96.631	377.3	91.1	95.346	91.507	220.4	58.8	97.023	94.326
ST U-Net	141.9	36.2	**98.168**	**96.466**	102.0	**27.6**	**98.544**	**97.141**	405.1	90.9	**95.789**	**92.196**	250.3	62.6	96.986	94.347
STR U-Net	**129.1**	**34.9**	97.945	96.141	**96.3**	28.0	98.258	96.710	**356.0**	**82.1**	95.603	92.002	**203.4**	**51.5**	**97.308**	**94.869**

Table 5
Ablation study on backbones and modules of the POST-IVUS framework, include the discriminator (Dis.) and TF module, based on EEM segmentation on NIRS-IVUS. The λ in discriminator Without specific description is set to 100.

	All				Normal				Side-branch				Calcium			
	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM
U-Net	164.3	42.3	97.369	95.096	124.2	33.2	97.743	95.767	435.9	102.0	94.690	90.408	251.7	65.0	96.584	93.563
U-Net + Dis. (λ = 10)	162.2	43.1	97.598	95.457	125.1	33.4	98.048	96.226	394.2	106.6	94.710	90.642	266.9	67.8	96.044	92.726
U-Net + Dis. (λ = 50)	163.4	42.5	97.132	94.626	129.2	34.6	97.526	95.329	395.1	94.2	94.381	89.832	242.0	62.4	96.060	92.639
U-Net + Dis. (λ = 100)	159.6	42.0	97.757	95.676	127.1	34.6	98.103	96.294	378.6	91.1	95.379	91.489	234.5	60.9	96.934	94.135
U-Net + TF	151.6	41.2	97.862	95.902	117.1	33.8	98.200	96.515	385.0	90.3	95.570	91.791	228.6	59.5	97.045	94.374
U-Net + Dis. + TF	143.8	37.6	97.992	96.123	112.6	30.6	98.331	96.731	349.3	82.7	95.695	92.064	223.6	56.9	97.136	94.516
ST U-Net	141.9	36.2	98.168	96.466	102.0	27.6	98.544	97.141	405.1	90.9	95.789	92.196	250.3	62.6	96.986	94.347
STR U-Net	129.1	34.9	97.945	96.141	96.3	28.0	98.258	96.710	356.0	82.1	95.603	92.002	203.4	51.5	97.308	94.869
ST U-Net + Dis. + TF	131.8	34.8	98.220	96.565	96.4	27.0	98.569	97.191	364.7	83.8	**96.013**	**92.599**	228.9	58.5	97.100	94.562
STR U-Net + Dis. + TF	**119.8**	**32.0**	**98.289**	**96.705**	**90.0**	**25.4**	**98.616**	**97.292**	**325.5**	**76.9**	95.966	92.596	**185.2**	**47.3**	**97.562**	**95.326**

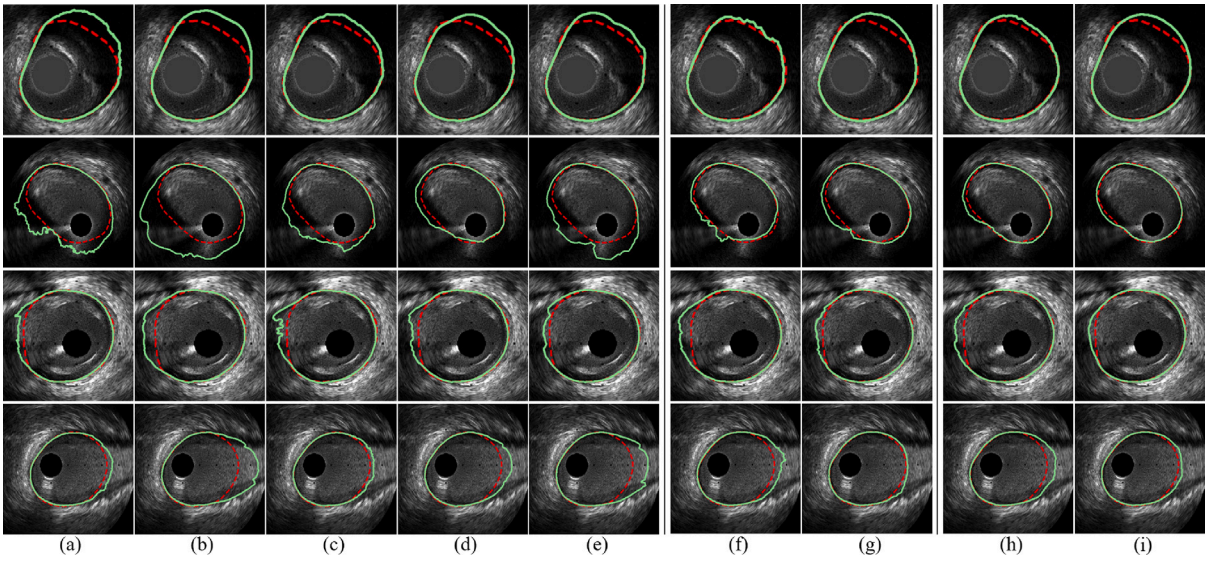


Fig. 8. A few EEM segmentation results in challenging areas, comparison between the proposed ST U-Net, STR U-Net and previous segmentation models on the NIRS-IVUS dataset: (a) SegNet (b) DeepLab V3+; (c) ENet; (d) GCN; (e) PSPNet; (f) ST U-Net; (g) STR U-Net; (h) ST U-Net + Dis. + TF; (i) STR U-Net + Dis. + TF. For all discriminators, $\lambda = 100$.

This combination of components ensures that STR U-Net addresses both the overall shape and intricate delicate boundary details in the segmentation task.

A comparative analysis is presented on the EEM segmentation task since it does not require temporal encoding and is more suitable for popular segmentation networks. Besides, the NIRS-IVUS EEM dataset provides a large number of stable annotations of EEM. Table 4 shows the results of several state-of-the-art networks on the IVUS segmentation task. We also develop ST U-Net, by replacing the Recurrent Residual convolutional unit by two convolutional layers to demonstrate the performance of RRCU in the segmentation branch. As it can be observed, the proposed ST U-Net and STR U-Net slightly improve the segmentation accuracy in normal class where no calcific plaque or side-branch is present. While in these tricky cases that require some inference ability, the two networks significantly improve the performance compared to the existing methods. As Hausdorff distance is considered the most important metric in this task, we uniformly use STR U-Net as the backbone in the following tasks due to its excellent performance in HD. The advantage of the proposed method can also be observed in Fig. 8, showing cases with shadow and side-branch regions, the proposed STR U-Net generates the most reasonable boundaries.

We also perform an ablation study on this task, results presented in Table 5. The aim is to reveal the impact of different modules on the performance of EEM segmentation. The POST-IVUS framework achieves superior segmentation accuracy due to its inference ability in side-branch and calcium cases. Specifically, the Temporal constraint and Fusion (TF) module slightly increase the overall accuracy, but due to its ability in suppressing errors in a single prediction, it significantly improves the performance in the side-branch and calcium cases. Compared to the original U-Net, the proposed ST U-Net and STR U-Net backbones improve the segmentation accuracy in all classes by a large margin. This improvement can also be observed from the visual results depicted in Fig. 10. Based on the original U-Net's results (a), adding the discriminator (b), the TF module (c)(d), and utilising the STR U-Net all bring improvements. Overall, the proposed scheme, namely, STR U-Net with the discriminator and TF module (e) achieve

Table 6
Lumen segmentation result for NIRS-IVUS dataset, comparison between different backbone, input data and encoders. For all discriminators, $\lambda = 100$.

	Encoder	All				Normal				Side-branch				Calcium			
		HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM	HD	MD	Dice	JM
U-Net	1 Frame	245.7	75.5	94.592	90.018	216.1	66.9	95.170	90.979	467.1	137.5	90.854	83.875	292.3	93.2	92.682	86.844
U-Net	3 Frame	220.2	68.9	95.265	91.231	191.5	60.7	95.869	92.224	435.7	129.9	91.170	84.639	274.4	93.0	93.261	88.013
U-Net	AE	217.8	69.6	95.237	91.242	186.6	58.3	95.942	92.354	447.6	147.6	91.187	84.758	303.7	120.4	92.134	86.678
U-Net	VE	215.3	65.2	95.244	91.241	187.3	57.4	95.931	92.347	426.9	121.9	90.928	84.301	253.3	79.4	92.955	87.625
U-Net + Dis.	3 Frame	215.7	64.0	95.505	91.584	190.5	57.6	95.982	92.388	418.7	114.5	92.116	85.935	237.0	71.0	94.427	89.710
U-Net + Dis.	AE	219.6	66.4	95.342	91.330	192.9	58.8	95.919	92.295	421.0	120.1	91.815	85.465	259.8	82.1	93.406	88.116
U-Net + Dis.	VE	213.7	62.8	95.565	91.707	185.5	55.3	96.121	92.639	430.3	118.5	91.877	85.585	246.7	74.7	94.028	89.104
U-Net + TF	3 Frame	197.6	59.3	95.800	92.125	170.9	52.6	96.333	93.017	406.0	110.6	92.047	85.953	225.8	68.5	94.628	90.050
U-Net + TF	AE	197.4	59.5	95.762	92.036	174.1	53.1	96.256	92.879	369.6	105.4	92.592	86.694	236.8	72.6	94.168	89.295
U-Net + TF	VE	196.6	58.9	95.850	92.212	169.4	51.8	96.376	93.099	411.9	112.4	92.260	86.231	228.6	69.5	94.454	89.821
U-Net + Dis. + TF	3 Frame	201.3	60.6	95.787	92.090	174.0	53.2	96.279	92.914	422.7	120.2	92.209	86.178	221.1	66.1	94.777	90.309
U-Net + Dis. + TF	AE	199.1	60.4	95.761	92.044	175.1	53.8	96.263	92.897	380.0	108.3	92.501	86.549	234.5	72.1	94.268	89.485
U-Net + Dis. + TF	VE	194.9	57.6	95.897	92.288	169.9	51.0	96.421	93.168	390.3	107.7	92.260	86.266	223.1	67.1	94.611	90.051
ST U-Net	3 Frame	184.1	56.9	96.094	92.635	156.0	50.1	96.586	93.476	408.9	109.8	92.603	86.764	206.2	64.1	95.054	90.762
ST U-Net	AE	192.8	59.4	95.922	92.310	165.0	52.8	96.401	93.138	405.2	109.3	92.727	86.798	225.9	68.3	94.669	90.145
ST U-Net	VE	173.6	53.5	96.294	92.959	147.3	47.2	96.717	93.698	378.6	100.9	93.334	87.844	202.8	61.7	95.240	91.099
STR U-Net	3 Frame	176.1	55.1	96.126	92.689	152.1	48.8	96.575	93.466	357.0	102.5	93.021	87.374	210.4	65.4	94.897	90.557
STR U-Net	AE	186.4	58.3	95.947	92.375	159.3	50.6	96.452	93.238	394.6	116.9	92.535	86.618	218.4	69.0	94.476	89.844
STR U-Net	VE	173.0	54.8	96.200	92.849	146.7	46.7	96.691	93.677	358.1	101.1	92.987	87.426	228.1	82.1	94.710	90.361
ST U-Net + Dis. + TF	3 Frame	172.2	52.8	96.317	93.032	145.4	46.6	96.790	93.840	386.4	101.7	92.865	87.192	194.9	58.0	95.466	91.510
ST U-Net + Dis. + TF	AE	182.4	56.7	96.060	92.563	156.3	50.2	96.537	93.384	387.4	106.9	92.795	86.956	206.5	63.4	94.967	90.678
ST U-Net + Dis. + TF	VE	165.5	50.5	96.446	93.250	141.6	44.9	96.873	93.984	359.3	95.9	**93.316**	**87.897**	**186.7**	**55.7**	**95.598**	**91.724**
STR U-Net + Dis. + TF	3 Frame	164.2	50.7	96.406	93.206	141.0	45.0	96.852	93.968	348.0	96.7	93.112	87.663	187.6	57.8	95.446	91.497
STR U-Net + Dis. + TF	AE	172.8	53.7	96.214	92.854	148.6	47.5	96.685	93.662	358.4	100.7	92.936	87.310	203.6	63.9	94.978	90.679
STR U-Net + Dis. + TF	VE	**162.2**	**49.9**	**96.447**	**93.262**	**139.4**	**44.2**	**96.888**	**94.021**	**341.7**	**94.2**	93.225	87.817	188.8	57.8	95.438	91.481

Table 7
Computer-to-observer difference and inter-observer agreement analysis.

	Lumen					EEM					Plaque area
	HD (μm)	MD (μm)	Dice (%)	JM (%)	Area (mm^2)	HD (μm)	MD (μm)	Dice (%)	JM (%)	Area (mm^2)	Area (mm^2)
Expert 1 - Expert 2	198.9 ± 147.2	58.5 ± 43.8	95.9 ± 3.0	92.3 ± 5.1	0.099 ± 0.534	133.8 ± 139.7	34.3 ± 35.9	98.2 ± 1.7	96.6 ± 2.9	−0.076 ± 0.660	0.174 ± 0.744
Expert 1 - Proposed	167.3 ± 128.8	51.8 ± 35.8	96.3 ± 3.0	93.0 ± 5.0	−0.035 ± 0.777	119.3 ± 151.3	31.6 ± 34.6	98.3 ± 2.1	96.8 ± 3.5	0.070 ± 0.617	−0.105 ± 0.697
Expert 2 - Proposed	169.2 ± 143.5	52.2 ± 37.3	96.3 ± 2.8	93.0 ± 4.7	0.064 ± 0.717	123.8 ± 139.7	34.3 ± 34.9	98.2 ± 1.7	96.6 ± 2.9	−0.005 ± 0.710	0.070 ± 0.516
modified Williams index	1.182	1.125	1.004	1.007	1.176	1.126	1.071	1.000	1.001	1.049	1.248
95% Confidence Interval	1.182, 1.183	1.124, 1.125	1.004, 1.004	1.007, 1.007	1.175, 1.177	1.125, 1.127	1.071, 1.072	1.000, 1.000	1.001, 1.001	1.047, 1.051	1.247, 1.248

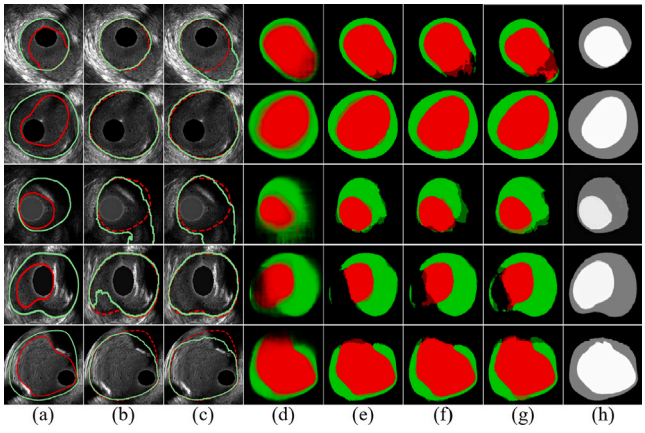


Fig. 9. Segmentation results based on different losses, on the NIRS-IVUS dataset: (a) ground truth, green: EEM border, red: lumen border; (b) $\mathcal{L}_{Boundary}$; (c) $\mathcal{L}_{Boundary} + \mathcal{L}_{Dice}$, green: prediction of EEM border, red: expert's annotation of EEM; (d) $\mathcal{L}_{CE}$; (e) $\mathcal{L}_{Dice}$; (f) $\mathcal{L}_{IOU}$; (g) $\mathcal{L}_{Tversky}$, green: prediction of EEM area, red: prediction of lumen area; (h) $\mathcal{L}_{L1}$, white representing lumen and grey EEM. All results are generated based on U-Net.

the best EEM boundaries in normal cases. Furthermore, the advantage of the proposed method is even more evident in calcific plaque and side-branch cases, where baseline methods are influenced in tricky regions while our method demonstrates excellent ability in inferring the boundaries when visual features are lacking or misleading.

4.4.3. NIRS-IVUS lumen segmentation

Lumen segmentation results are presented in Table 6. As it can be observed, the use of discriminator and TF modules also improves performance, similar to the EEM segmentation results. The lumen segmentation task is more challenging due to the insufficient features in static frames. The TF module can eliminate some of the errors by considering the temporal context, thus providing the most performance gains in lumen segmentation.

Furthermore, experiments demonstrate that lumen boundaries cannot be predicted accurately based on only one frame. This aligns with our rationale for exploiting temporal features from neighbouring frames for lumen segmentation. We test the two proposed encoders and use three adjacent frame inputs as controls. Based on Table 6, the alignment encoder (AE) performs slightly inferiorly to the approach that combines three frames together (3 Frames). This could be because the rotation of the original image introduces errors. The visual persistence encoder (VE) is designed to mimic the annotation method of human experts and effectively capture motion information. This design contributes to enhanced lumen segmentation results compared to other methods.

The same performance can be observed in Fig. 11. Comparing results in (b) to (a), it can be seen that including EEM annotation in training improves the boundary quality of lumen. Further enhancement can be observed in (c) and (d) when three consecutive frames are added for feature extraction. In (e) and (f), the alignment encoder is applied together with the STR U-Net, with or without the discriminator and the TF module. Finally, (g) and (h) shows the result when the visual persistence encoder is involved in the process, with STR U-Net, with and without discriminator and the TF module. Overall, the proposed method, i.e., the STR U-Net with discriminator, the TF module and the visual persistence encoder (h), achieves the best lumen boundaries that well resemble human annotation.

4.4.4. Computer-to-observer agreement and inter-observer agreement

To evaluate the agreement between the computer model and the observers, as well as the agreement between the observers themselves, two experts independently annotate the test set. Modified William index is calculated for every pair of results between expert 1 and expert 2, expert 1 and the proposed model, and expert 2 and the proposed model. According to the results reported in Table 7, the mWI of the

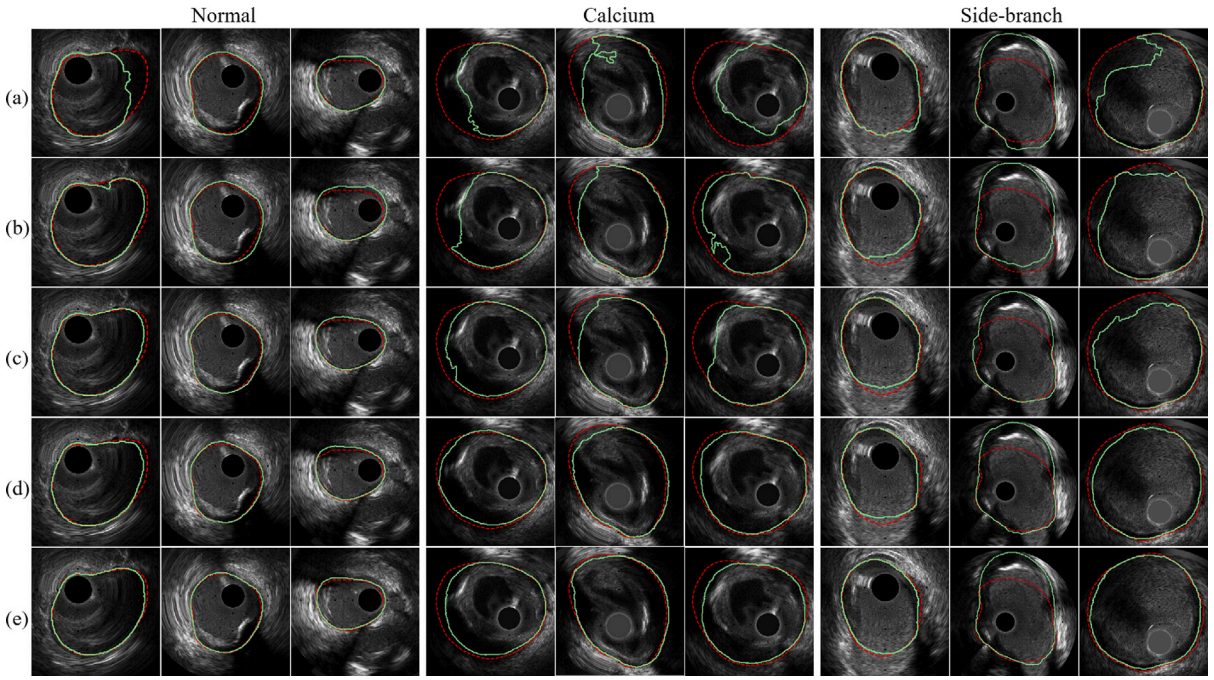


Fig. 10. Some examples of EEM segmentation results on the NIRS-IVUS dataset: (a) U-Net; (b) U-Net + discriminator, $\lambda = 100$; (c) U-Net + TF; (d) U-Net + discriminator + TF, $\lambda = 100$; (e) STR U-Net + discriminator + TF, $\lambda = 100$. Light green solid lines indicate machine annotation, red dash lines represent human expert 1's annotation.

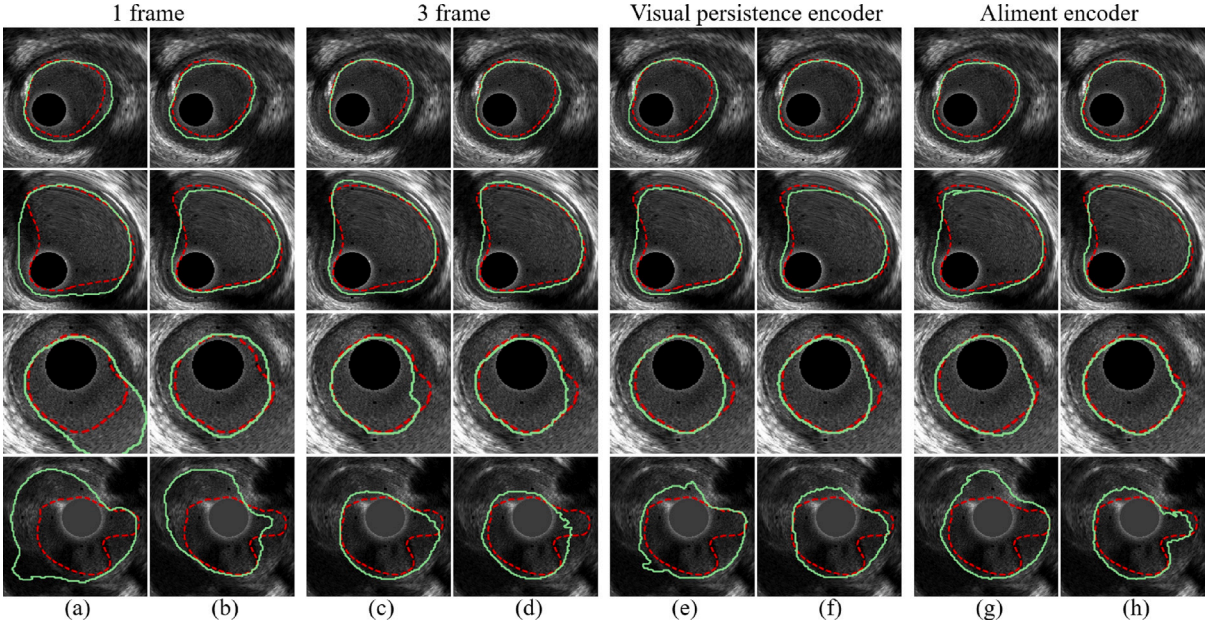


Fig. 11. Lumen segmentation results on NIRS-IVUS dataset. Red lines: ground truth; green dash line: predicted boundaries. (a) single frame input, only lumen annotation (STR U-Net); (b) single frame input, both lumen and EEM annotation (STR U-Net); (c) with 3-frames input (STR U-Net); (d) with 3-frames input (STR U-Net + Dis. + TF); (e) with alignment encoder (STR U-Net); (f) with alignment encoder (STR U-Net + Dis. + TF); (g) with visual persistence encoder (STR U-Net); (h) with visual persistence encoder (STR U-Net + Dis. + TF). For all discriminators, $\lambda = 100$.

proposed method is greater than 1 in all cases, which indicates that the predicted boundaries by the model are highly likely to be consistent with the manual annotation. The width of the 95% confidence interval, representing the sample variability, is small, and both the lower and upper limits of the interval are above 1 in all cases. The level of computer-to-observer agreement is greater than inter-observer agreement, meaning the model's predicted boundaries represent the consensus of the two experts. The advantages of the proposed method are more evident in the lumen segmentation task. This indicates that the proposed perceptual organisation based inference and encoding schemes are able to produce boundaries that highly resemble human annotations that are sketched by utilising their semantic understanding and expertise knowledge.

More specifically, the computer-to-observer and inter-observer variability distributions in terms of Hausdorff distance are illustrated in Fig. 12. The green bars demonstrate that the variances between the model results and expert annotations is smaller compared to that between the two experts. The pink violin plots show that the major distribution of the variability between the model results and expert

Table 8
A comparison of results of the proposed method and other reported methods since 2018 on the IVUS-2011 dataset B, in the three metrics utilised by the benchmark. The Hausdorff distance is in millimetres.

			Yang et al. (2018)	Faraji et al. (2018)	Hammouche et al. (2019)	Kermani et al. (2019)	Lo Vercio et al. (2019)	Yang et al. (2019)	Huang et al. (2020)	Gao et al. (2020)	Xia et al. (2020)	Szarski et al. (2021)	Huang et al. (2021)	Blanco et al. (2022)	Proposed
All	Lumen	HD	0.26 (0.25)	0.30 (0.20)	0.31 (0.16)	0.38 (0.26)	0.32 (0.25)	0.25 (0.20)	0.28 (0.17)	0.22 (0.15)	0.27 (0.13)	0.29 (0.22)	0.30 (0.25)	0.241 (0.164)	**0.21 (0.13)**
		JM	0.90 (0.06)	0.87 (0.06)	0.90 (0.05)	0.84 (0.07)	0.88 (0.08)	0.90 (0.05)	0.89 (0.05)	**0.92**	0.90 (0.04)	0.90 (0.08)	0.87 (0.06)	0.902 (0.049)	**0.92 (0.04)**
		Pad	–	0.08 (0.09)	0.06 (0.06)	0.10 (0.08)	0.09 (0.10)	–	0.10 (0.08)	–	0.06 (0.06)	–	0.10 (0.11)	0.058 (0.058)	**0.05 (0.04)**
	EEM	HD	0.48 (0.44)	0.67 (0.54)	–	0.64 (0.41)	0.57 (0.31)	0.30 (0.35)	0.47 (0.31)	0.29 (0.22)	0.37 (0.31)	0.32 (0.30)	0.37 (0.20)	0.205 (0.152)	**0.19 (0.18)**
		JM	0.86 (0.11)	0.77 (0.17)	–	0.82 (0.11)	0.83 (0.10)	0.92 (0.07)	0.84 (0.11)	0.92	0.90 (0.07)	0.91 (0.09)	0.85 (0.08)	0.930 (0.041)	**0.94 (0.05)**
		Pad	–	0.19 (0.18)	–	0.13 (0.11)	0.13 (0.09)	–	0.13 (0.13)	–	0.06 (0.06)	–	0.11 (0.09)	0.041 (0.048)	**0.04 (0.05)**
No Artefact	Lumen	HD	0.21 (0.09)	0.29 (0.17)	0.26 (0.11)	0.36 (0.21)	0.29 (0.16)	0.25 (0.17)	–	–	–	–	–	–	**0.20 (0.14)**
		JM	0.91 (0.03)	0.88 (0.05)	0.90 (0.04)	0.85 (0.07)	0.89 (0.05)	0.91 (0.04)	–	–	–	–	–	–	**0.92 (0.04)**
		Pad	–	0.08 (0.07)	**0.05 (0.04)**	0.10 (0.08)	0.07 (0.07)	–	–	–	–	–	–	–	**0.05 (0.04)**
	EEM	HD	0.27 (0.23)	0.31 (0.23)	–	0.43 (0.23)	0.38 (0.24)	0.17 (0.08)	–	–	–	–	–	–	**0.15 (0.10)**
		JM	0.92 (0.05)	0.89 (0.07)	–	0.87 (0.05)	0.89 (0.06)	**0.95 (0.02)**	–	–	–	–	–	–	**0.95 (0.03)**
		Pad	–	0.07 (0.08)	–	0.11 (0.06)	0.06 (0.05)	–	–	–	–	–	–	–	**0.02 (0.03)**
Bifurcation	Lumen	HD	0.50 (0.58)	0.53 (0.34)	0.40 (0.21)	0.47 (0.32)	0.44 (0.33)	0.46 (0.38)	0.48 (0.33)	–	–	–	0.37 (0.17)	–	**0.31 (0.18)**
		JM	0.82 (0.11)	0.79 (0.10)	0.85 (0.07)	0.83 (0.07)	0.84 (0.09)	0.85 (0.10)	0.83 (0.10)	–	–	–	0.85 (0.04)	–	**0.89 (0.06)**
		Pad	–	0.15 (0.17)	**0.08 (0.10)**	0.12 (0.08)	0.12 (0.13)	–	0.17 (0.19)	–	–	–	0.11 (0.07)	–	**0.08 (0.06)**
	EEM	HD	0.82 (0.60)	1.22 (0.45)	–	0.99 (0.53)	0.68 (0.34)	0.60 (0.35)	0.58 (0.36)	–	–	–	0.49 (0.17)	–	**0.32 (0.23)**
		JM	0.78 (0.11)	0.57 (0.13)	–	0.74 (0.13)	0.79 (0.12)	0.86 (0.10)	0.81 (0.14)	–	–	–	0.83 (0.07)	–	**0.92 (0.05)**
		Pad	–	0.32 (0.19)	–	0.22 (0.20)	0.15 (0.11)	–	0.17 (0.17)	–	–	–	0.11 (0.06)	–	**0.06 (0.07)**
Side Vessels	Lumen	HD	0.23 (0.12)	0.24 (0.11)	0.25 (0.12)	0.34 (0.32)	0.31 (0.27)	0.20 (0.12)	0.36 (0.36)	–	–	–	0.24 (0.24)	–	**0.19 (0.10)**
		JM	0.90 (0.04)	0.87 (0.05)	0.88 (0.05)	0.85 (0.08)	0.88 (0.09)	0.91 (0.04)	0.87 (0.11)	–	–	–	0.89 (0.05)	–	**0.92 (0.03)**
		Pad	–	0.06 (0.05)	0.05 (0.04)	0.11 (0.09)	0.09 (0.11)	–	0.15 (0.13)	–	–	–	0.08 (0.07)	–	**0.04 (0.03)**
	EEM	HD	0.59 (0.49)	0.74 (0.18)	–	0.77 (0.46)	0.61 (0.29)	0.35 (0.36)	0.52 (0.31)	–	–	–	0.44 (0.24)	–	**0.23 (0.23)**
		JM	0.83 (0.14)	0.73 (0.60)	–	0.77 (0.12)	0.81 (0.11)	0.91 (0.08)	0.81 (0.11)	–	–	–	0.84 (0.06)	–	**0.93 (0.06)**
		Pad	–	0.21 (0.18)	–	0.16 (0.13)	0.14 (0.09)	–	0.18 (0.18)	–	–	–	0.13 (0.09)	–	**0.05 (0.06)**
Shadow	Lumen	HD	0.27 (0.25)	0.29 (0.20)	0.28 (0.13)	0.36 (0.22)	0.28 (0.19)	0.25 (0.20)	0.29 (0.26)	–	–	–	–	–	**0.20 (0.11)**
		JM	0.87 (0.06)	0.86 (0.07)	0.86 (0.07)	0.83 (0.07)	0.87 (0.06)	0.89 (0.05)	0.88 (0.08)	–	–	–	–	–	**0.91 (0.05)**
		Pad	–	0.08 (0.09)	0.06 (0.06)	0.11 (0.07)	0.07 (0.07)	–	0.12 (0.14)	–	–	–	–	–	**0.05 (0.05)**
	EEM	HD	0.80 (0.45)	1.24 (0.39)	–	1.01 (0.39)	0.67 (0.36)	0.48 (0.48)	0.56 (0.33)	–	–	–	–	–	**0.27 (0.24)**
		JM	0.76 (0.12)	0.58 (0.13)	–	0.72 (0.12)	0.77 (0.12)	0.88 (0.10)	0.76 (0.13)	–	–	–	–	–	**0.92 (0.06)**
		Pad	–	0.37 (0.15)	–	0.12 (0.13)	0.16 (0.11)	–	0.23 (0.19)	–	–	–	–	–	**0.06 (0.07)**

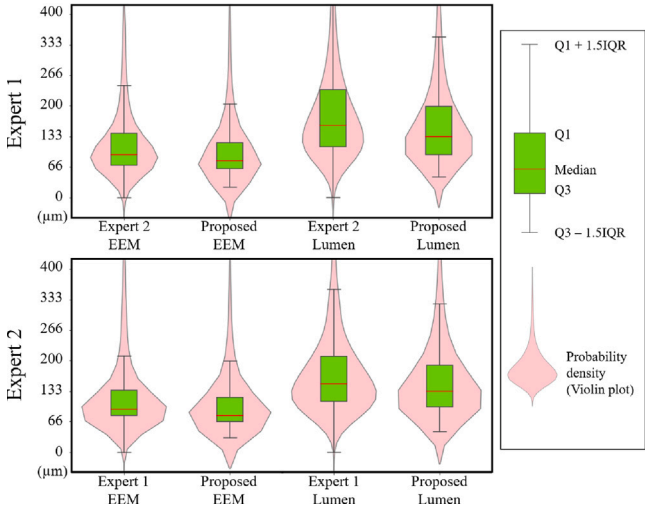


Fig. 12. The violin plot and box plot of Hausdorff distance. Compare all three annotations with each other. Q1 and Q3 are also denote as 25th percentile and 75th percentile. The IQR can be calculated by $Q1 - Q3$.

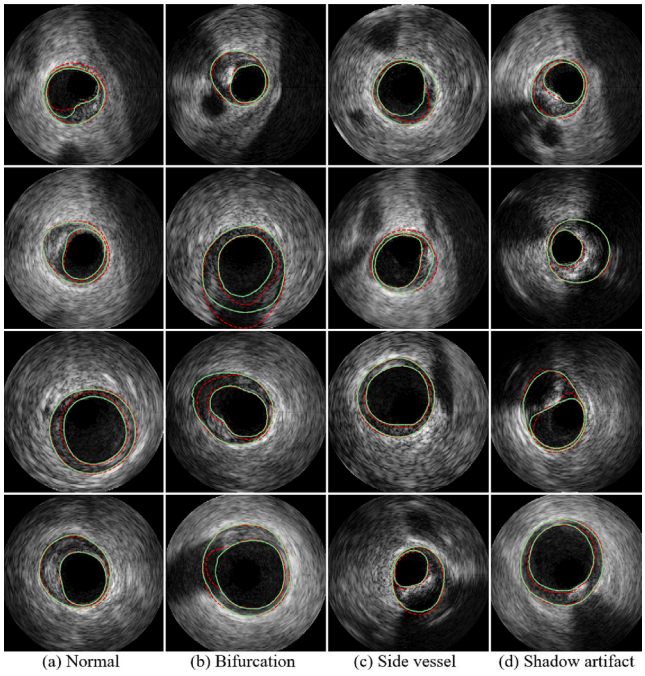


Fig. 13. Some segmentation results from IVUS-2011 dataset. Light green solid lines: machine annotation, red dash lines: expert annotation.

annotations is smaller (lower in values) than that between the two experts.

Experts' visual assessment on representative samples also agrees with this conclusion. Our models are integrated into QCU-CMS software and used by cardiology teams at multiple institutions in their daily work. According to the experts' feedback, they find the predicted boundaries by the POST-IVUS platform highly agreeable, compared to other observers' annotations, and little modification to the predicted boundaries is needed before it can be considered the gold standard.

4.4.5. Results on IVUS-2011 dataset

The POST-IVUS framework proposed in this paper is also evaluated on the open IVUS-2011 dataset of coronary artery images (dataset B, 20MHz), in which 109 frames are used as the training set, and the remaining 326 frames are used as the test set. This experiment allows fair performance comparison against the existing state-of-the-art methods. We summarise the evaluation results of POST-IVUS as well as the major methods reported since 2018 in Table 8.

The table shows that the proposed POST-IVUS segmentation framework achieves a new state-of-the-art performance on this task, with the best scores in almost all evaluation metrics. The class with 'no

Table 9
Comparison of network size, training time, and evaluation time for various segmentation methods, including the proposed approach (Encoder + STR U-Net + Dis. + TF). All the time evaluations are based on a single image, and the time unit is seconds.

Method	Network size	Training time	Evaluating time
U-Net (Ronneberger et al., 2015)	8M	0.011 s	<0.1 s
SegNet (Badrinarayanan et al., 2017)	16M	0.020 s	<0.1 s
PSPNet (Zhao et al., 2017)	84M	0.054 s	<0.1 s
DeepLab V3+ (Chen et al., 2018)	54M	0.022 s	<0.1 s
ENet (Paszke et al., 2016)	0.4M	0.013 s	<0.1 s
GCN (Kipf and Welling, 2016)	23M	0.013 s	<0.1 s
ST U-Net	41M	0.089 s	<0.1 s
STR U-Net	42M	0.090 s	<0.1 s
STR U-Net+Dis.	42M	0.090 s	<0.1 s
STR U-Net+Dis.+TF	42M	0.091 s	0.45 s
Encoder+STR U-Net+Dis.+TF	42M	0.091 s	0.45 s

artifacts' is relatively easy to address with a high segmentation accuracy on the pixel level, and thus many traditional segmentation networks can achieve the same high scores in some metrics. Overall, POST-IVUS achieves the best scores with a small margin. However, in more challenging cases of bifurcation, side vessels, and shadow classes, POST-IVUS shows an evident advantage due its inference ability in tricky regions. Superior results are scored in all three metrics considered in this dataset.

Some visual results are shown in Fig. 13. In IVUS segmentation tasks, the correctness of the boundaries is considered as the most important indicator instead of region overlap. This is the reason why Hausdorff distance is often used as the main metric. As it can be observed in this figure, the predicted boundaries by POST-IVUS resemble the annotated boundaries even in dark areas with no obvious visual features available. This is the main added value of the proposed method on top of the existing pixel-level segmentation methods.

4.5. Network size and time analysis

In this section, we provide a comparative analysis of the network size, training time, and evaluation time for various methods, including our proposed approach, as shown in Table 9:

Based on Table 9, we analyse the relationship between network size, training time, and evaluation time. The proposed method (Encoder + STR U-Net + Dis. + TF) has a network size of 42M, which is relatively moderate compared to some other methods like PSPNet and DeepLab V3+. Although the Transformer-based methods, i.e., in rows 7 to 11, exhibit improved performance and have smaller Network sizes, their training efficiency is lower compared to other models. This is primarily due to the self-attention mechanism employed in Transformers, which increases the computational complexity and memory requirements, ultimately resulting in longer training times. The inclusion of the Discriminator during training does not lead to a significant increase in training time due to its simple structure, which does not impose a substantial burden on the GPU. Additionally, the proposed Encoders do not considerably affect runtime as they are computationally efficient.

In this task, evaluation time is more critical than training time. While an extended training period may be acceptable, once deployed, hospitals will need to analyse blood vessels with limited computational resources, making the evaluation time more crucial. For evaluating time, the model itself is not the primary bottleneck, as it only involves forward propagation, which is relatively fast. The most significant bottleneck lies in the post-processing computation, especially the TF module. It requires numerous complex evaluations and fusion strategies, leading to an extended evaluation time of around 0.45 s per frame. However, compared to the accuracy gains provided by post-processing, the additional time spent is considered worthwhile.

5. Clinical impact

We have integrated the proposed POST-IVUS framework into the QCU-CMS software and applied it to pullbacks generated by several types of IVUS catheters. This advances have enabled fast analysis of large datasets and more reproducible evaluation of atheroma burden in longitudinal studies. For a typical vessel, the average segmentation time is reduced from about 10 h to 10 min. Due to the dedicated designs in the POST-IVUS framework that imitate the perceptual organisation ability of human vision, the resulting boundaries achieve superior segmentation accuracy compared to previously proposed methods. The boundaries are very close to the manually annotated ones, and are found even more agreeable by multiple experts than their estimations. Furthermore, the robustness of the POST-IVUS framework ensures that most common errors are eliminated, and little editing efforts are required to finalise analysis.

Based on the inter-observer variability evaluation between the model and expert results, the model surpasses the performance of expert annotations in some areas, especially when demarcating the lumen border. The output of POST-IVUS saves time and resources for the clinical and research teams, improves the diagnosis procedures and has a potential for clinical translation.

6. Conclusion

This paper presented a universal POST-IVUS segmentation framework based on EEM and lumen borders. This framework includes a dedicated set of temporal context-based feature encoders for extracting descriptive temporal features for lumen boundary, a selective transformer recurrent U-Net that achieves high-resolution segmentation, inference-based segmentation on challenging areas, an adversarial learning scheme to ensure that the boundary inference is aware of the perceptual organisation property of human vision, and a temporal constraint and fusion module to further eliminate errors and improve the robustness of results. The POST-IVUS framework is superior to previous methods in segmenting EEM and lumen areas under challenging conditions like when calcification, side-branch, and other artifacts are present. According to the estimation of model-observer and inter-observer variability, the model-predicted boundaries reached a high consensus with the observers that was superior to the agreement of the experts. The method has been integrated into the QCU-CMS software to provide cardiologists with high-precision EEM and lumen segmentation which form a good foundation for their subsequent IVUS-based analysis. We are working jointly with additional vascular imaging modalities such as IVUS-OCT and histology to further expand plaque segmentation and analysis in the future.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christos Bourantas reports financial support was provided by British Heart Foundation. Qianni Zhang, Christos Bourantas has patent #sn: 2003871.1 issued to Queen Mary University of London.

Data availability

The authors do not have permission to share data.

References

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bajaj, R., Huang, X., Kilic, Y., Ramasamy, A., Jain, A., Ozkor, M., Tufaro, V., Safi, H., Erdogan, E., Serruys, P.W., et al., 2021. Advanced deep learning methodology for accurate, real-time segmentation of high-resolution intravascular ultrasound images. Int. J. Cardiol. 339, 185–191.

Balocco, S., Basset, O., Cachard, C., Delachartre, P., 2003. Spatial anisotropic diffusion and local time correlation applied to segmentation of vessels in ultrasound image sequences. In: IEEE Symposium on Ultrasonics, Vol. 2. IEEE, pp. 1549–1552.

Balocco, S., Gatta, C., Ciompi, F., Wahle, A., Radeva, P., Carlier, S., Unal, G., Sanidas, E., Mauri, J., Carillo, X., et al., 2014. Standardized evaluation methodology and reference database for evaluating IVUS image segmentation. Comput. Med. Imaging Graph. 38 (2), 70–90.

Bargsten, L., Raschka, S., Schlaefer, A., 2021. Capsule networks for segmentation of small intravascular ultrasound image datasets. Int. J. Comput. Assist. Radiol. Surg. 1–12.

Blanco, P.J., Ziemer, P.G., Bulant, C.A., Ueki, Y., Bass, R., Räber, L., Lemos, P.A., García-García, H.M., 2022. Fully automated lumen and vessel contour segmentation in intravascular ultrasound datasets. Med. Image Anal. 75, 102262.

Bourantas, C.V., Plissiti, M.E., Fotiadis, D.I., Protopappas, V.C., Michalis, L.K., 2005. In vivo validation of a novel automated method for border detection in IVUS images. Br. J. Radiol. 78 (926), 122–129.

Chalana, V., Kim, Y., 1997. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans. Med. Imaging 16 (5), 642–652.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Cui, W., Zeng, L., Chong, B., Zhang, Q., 2021. Toothpix: pixel-level tooth segmentation in panoramic X-Ray images based on generative adversarial networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1346–1350.

Dai, B., Fidler, S., Urtasun, R., Lin, D., 2017. Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2970–2979.

Destrempes, F., Cardinal, M.R., Allard, L., Tardif, J.C., Cloutier, G., 2014. Segmentation method of intravascular ultrasound images of human coronary arteries. Comput. Med. Imaging Graph. Off. J. Comput. Med. Imaging Soc. 38 (2), 91–103.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Essa, E., Xie, X., Sazonov, I., Nithiarasu, P., SmIth, D., 2012. Shape prior model for media-adventitia border segmentation in ivUS using graph cut. In: MICCAI Medical Computer Vision.

Faraji, M., Cheng, I., Naudin, I., Basu, A., 2018. Segmentation of arterial walls in intravascular ultrasound cross-sectional images using extremal region selection. Ultrasonics 84, 356–365.

Gao, Z., Chung, J., Abdelrazek, M., Leung, S., Hau, W.K., Xian, Z., Zhang, H., Li, S., 2020. Privileged modality distillation for vessel border detection in intracoronary imaging. IEEE Trans. Med. Imaging 39 (5), 1524–1534.

Gao, Y., Zhou, M., Metaxas, D.N., 2021. UTNet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 61–71.

Gil, D., Hernandez, A., Rodriguez, O., Mauri, J., Radeva, P., 2006. Statistical strategy for anisotropic adventitia modelling in IVUS. IEEE Trans. Med. Imaging 25 (6), 768–778.

Gil, D., Radeva, P., Saludes, J., Mauri, J., 2001. Automatic Segmentation of Artery Wall in Coronary IVUS Images: A Probabilistic Approach. IEEE.

Ginestar, D., Hueso, J.L., Riera, J., Lázaro, I., 2014. Semi-automatic segmentation of IVUS images for the diagnosis of cardiac allograft vasculopathy.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Hammouche, A., Cloutier, G., Tardif, J.C., Hammouche, K., Meunier, J., 2019. Automatic IVUS lumen segmentation using a 3D adaptive helix model. Comput. Biol. Med. 107, 58–72.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Huang, Y., Xia, M., Guo, Y., Zhou, G., Wang, Y., 2021. Extraction of media adventitia and luminal intima borders by reconstructing intravascular ultrasound image sequences with vascular structural continuity. Med. Phys. 48 (8), 4350–4364.

Huang, Y., Yan, W., Xia, M., Guo, Y., Zhou, G., Wang, Y., 2020. Vessel membrane segmentation and calcification location in intravascular ultrasound images using a region detector and an effective selection strategy. Comput. Methods Programs Biomed. 189, 105339.

Iskurt, A., CanDeMir, S., Akgul, Y.S., 2006. Identification of luminal and medial adventitial borders in intravascular ultrasound images using level sets. In: International Symposium on Computer and Information Sciences.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M.T., Kavousi, M., Stevens, G., Riley, L.M., Savin, S., Khan, T., Altay, S., et al., 2019. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob. Health 7 (10), e1332–e1345.

Katouzian, A., Angelini, E.D., Sturm, B., Laine, A.F., 2010. Automatic detection of luminal borders in IVUS images by magnitude-phase histograms of complex brushlet coefficients. In: Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, Vol. 2010, No. 1. pp. 3073–3076.

Kermani, A., Ayatollahi, A., 2019. A new nonparametric statistical approach to detect lumen and media-adventitia borders in intravascular ultrasound frames. Comput. Biol. Med. 104, 10–28.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B., 2019. Boundary loss for highly unbalanced segmentation. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 285–296.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S., 2020. Skin lesion segmentation via generative adversarial networks with dual discriminators. Med. Image Anal. 64, 101716.

Li, Y.C., Shen, T.Y., Chen, C.C., Chang, W.T., Lee, P.Y., Huang, C.C.J., 2021. Automatic detection of atherosclerotic plaque and calcification from intravascular ultrasound images by using deep convolutional neural networks. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 68 (5), 1762–1772.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 510–519.

Li, Y., Xu, Z., Wang, Y., Zhou, H., Zhang, Q., 2020. Su-net and du-net fusion for tumour segmentation in histopathology images. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 461–465.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Mendizabalruiz, G., Rivera, M., Ioannis, A., Mendizabalruiz, G., Rivera, M., Kakadiaris, I.A., 2010. (CC/CIMAT) A probabilistic segmentation method for IVUS images.

Mintz, G.S., Nissen, S.E., Anderson, W.D., Bailey, S.R., Erbel, R., Fitzgerald, P.J., Pinto, F.J., Rosenfield, K., Siegel, R.J., Tuzcu, E.M., et al., 2001. American College of Cardiology clinical expert consensus document on standards for acquisition, measurement and reporting of intravascular ultrasound studies (ivus). A report of the american college of cardiology task force on clinical expert consensus documents developed in collaboration with the european society of cardiology endorsed by the society of cardiac angiography and interventions. Eur. J. Echocardiogr. 2 (4), 299–313.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Nie, D., Shen, D., 2020. Adversarial confidence learning for medical image segmentation and synthesis. Int. J. Comput. Vis. 128 (10), 2494–2513.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier gans. In: International Conference on Machine Learning. PMLR, pp. 2642–2651.

O'malley, S.M., Naghavi, M., Kakadiaris, I.A., 2007. One-class acoustic characterization applied to blood detection in IVUS. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 202–209.

Paszke, A., Chaurasia, A., Kim, S., Culurciello, E., 2016. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rosales, M., Radeva, P., Rodriguez-Leor, O., Gil, D., 2009. Modelling of image-catheter motion for 3-D IVUS. Med. Image Anal. 13 (1), 91–104.

Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J., Benziger, C.P., et al., 2020. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J. Am. Coll. Cardiol. 76 (25), 2982–3021.

Sheet, D., Karamalis, A., Eslami, A., Noël, P., Chatterjee, J., Ray, A.K., Laine, A.F., Carlier, S.G., Navab, N., Katouzian, A., 2014. Joint learning of ultrasonic backscattering statistical physics and signal confidence primal for characterizing atherosclerotic plaques using intravascular ultrasound. Med. Image Anal. 18 (1), 103–117.

Sonka, M., Zhang, X., 1995. Segmentation of intravascular ultrasound images: a knowledge-based approach. IEEE Trans. Med. Imaging 14 (4), 719–732.
Sun, Y., Huang, X., Zhou, H., Zhang, Q., 2021. SRPN: similarity-based region proposal networks for nuclei and cells detection in histology images. Med. Image Anal. 102142.
Szarski, M., Chauhan, S., 2021. Improved real-time segmentation of intravascular ultrasound images using coordinate-aware fully convolutional networks. Comput. Med. Imaging Graph. 91, 101955.
Taki, A., Najafi, Z., Roodaki, A., Setarehdan, S.K., Zoroofi, R.A., Konig, A., Navab, N., 2008. Automatic segmentation of calcified plaques and vessel borders in IVUS images. Int. J. Comput. Assist. Radiol. Surg. 3 (3), 347–354.
Tufaro, V., Serruys, P.W., Räber, L., Bennett, M.R., Torii, R., Gu, S.Z., Onuma, Y., Mathur, A., Baumbach, A., Bourantas, C.V., 2023. Intravascular imaging assessment of pharmacotherapies targeting atherosclerosis: advantages and limitations in predicting their prognostic implications. Cardiovasc. Res. 119 (1), 121–135.
Ünal, G., Bucher, S., Carlier, S., Slabaugh, G., Fang, T., Tanaka, K., 2008. Shape-driven segmentation of the arterial wall in intravascular ultrasound images. IEEE Trans. Inf. Technol. Biomed. 12 (3), 335–347.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Vercio, L.L., del Fresno, M., Larrabide, I., 2019. Lumen-intima and media-adventitia segmentation in IVUS images using supervised classifications of arterial layers and morphological structures. Comput. Methods Programs Biomed. 177, 113–121.
Xia, M., Yan, W., Huang, Y., Guo, Y., Zhou, G., Wang, Y., 2020. Extracting membrane borders in ivus images using a multi-scale feature aggregated u-net. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, IEEE, pp. 1650–1653.
Yang, J., Faraji, M., Basu, A., 2019. Robust segmentation of arterial walls in intravascular ultrasound images using Dual Path U-net. Ultrasonics 96, 24–33.
Yang, J., Tong, L., Faraji, M., Basu, A., 2018. IVUS-net: An intravascular ultrasound segmentation network. In: International Conference on Smart Multimedia.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.