scientific reports

OPEN

Seg2RefineNet: a novel DL-based framework for 2D CCTA image-based segmentation and 3D volume-based refinement

Umair Khan & Panos Liatsis

Computed Tomography Coronary Angiography is a non-invasive imaging technique widely used to assess structural abnormalities, blockages, or narrowing (stenosis) of coronary arteries, thereby aiding in the diagnosis and management of coronary heart disease. To assist clinicians in the assessment process, various AI-based methods have been proposed, both for 2D and 3D data, to accurately extract / segment the coronary arterial tree. This work aims to develop a novel two-stage hybrid segmentation method, Seg2RefineNet, to enhance coronary artery segmentation. The first stage employs a 2D spatio-frequency attention UNet, which results in the initial segmentation providing precise vessel boundary identification with high resolution. The second stage refines the segmentation using a 3D Attention-GAN, incorporating the inter-slice relationships within the 3D volume. As a proof-of-concept, this novel DL-based framework is evaluated on the largest publicly available dataset ImageCAS, outperforming the existing state-of-the-art methods by achieving a mean Dice score of 0.8313 and a Hausdorff distance of 12.95 mm. This hybrid approach effectively combines the strengths of both 2D and 3D models, setting a new benchmark for coronary artery segmentation.

Cardiovascular Disease (CVD) represents one of the most pressing global health challenges, with the World Heart Federation reporting approximately 20.5 million deaths worldwide in 2021, accounting for one-third of all deaths[1]. Coronary Heart Disease (CHD) is responsible for coronary vessel narrowing due to calcium and fatty deposits within the arterial walls[2], leading to potentially blood flow-limiting stenosis, which causes decreased myocardial perfusion and ultimately, myocardial infarction. Accurate quantification of coronary artery stenosis is therefore essential for CHD patient risk assessment and treatment planning.

Coronary Computed Tomography Angiography (CCTA) is the gold standard non-invasive imaging technique, providing high-resolution 3D visualization of the coronary arterial tree for diagnosis and treatment planning. Typically, radiologists have to manually locate the coronary arteries, isolate their boundaries, and quantitatively analyze regions of stenosis. This is an inherently time-consuming process, which is prone to inter-observer variability, and increasingly challenging, given the growing volume and complexity of medical imaging data. To address these limitations, a number of deep learning (DL)-based automated segmentation methods have been proposed, broadly categorized into three approaches, i.e., 2D slice-based methods, which leverage high-resolution spatial information, 3D volumetric techniques that capture global contextual relationships, and hybrid frameworks that combine the advantages of both of the aforementioned approaches.

Jia et al. provided a systematic review of automated coronary segmentation research, highlighting that they show significant promise in addressing the scalability and accuracy challenges of manual segmentation[3]. 2D slice-based segmentation techniques are capable of providing high-resolution results with enhanced boundary precision. In this regard, Cheung et al. proposed an encoder-decoder architecture using transpose convolutions to restore spatial information during decoding[4]. Hong et al. introduced dual attention coordination mechanisms utilizing multi-level spatial attention to highlight vessel-related spatial features[5]. To address the annotation challenge in large-scale datasets, Chen et al. utilized positive unlabeled learning, demonstrating significant improvements over fully-supervised approaches[6]. Fu et al. showed that 2D methods can achieve superior boundary delineation, however, they struggle in integrating inter-slice contextual information[7].

Moving to volumetric approaches, fully 3D methods leverage complete spatial context to capture complex vessel relationships. Lei et al. proposed a 3D fully convolutional network with attention gates for end-to-end binary segmentation[8]. Shen et al. improved the approach by incorporating level set-based optimization for segmentation refinement[9].Wang et al. proposed attention-guided mechanisms for joint coronary artery and

Department of Computer Science, Khalifa University, Abu Dhabi 20000, UAE. email: panos.liatsis@ku.ac.ae

vein segmentation with topological consistency[10], while Liu et al. developed the Attention Guided and Feature Aggregated Network (AGFA-Net), which included multi-level feature enhancement through channel and spatial attention, combined with dilated convolutions[11]. A key coronary segmentation challenge is the preservation of topological structure, and ensuring anatomical connectivity in the entire vascular tree. Qiu et al. tackled this by a three-stage topology preservation framework aimed at fully connected coronary artery extraction and focused on centerline connectivity and branch topology preservation[12]. Zhang et al. proposed the use of topology-aware loss functions, which supported structural consistency during segmentation[13]. Kong et al. targeted topological modeling by using tree-structured convolutional gated recurrent units (ConvGRU) to capture the coronary anatomy, integrating voxel-based features with topological relationships from the ConvGRU[14]. 3D approaches suffer from inherent computational complexity, which led to the development of novel processing strategies. Chen et al. applied patch-based approaches, selecting $32 \times 32 \times 32$ volumes of interest as input to 3D UNet, combined with Frangi vessel enhancement for multi-channel processing[15]. Huang et al. demonstrated that larger patch sizes yield superior performance compared to smaller alternatives[16]. Zeng et al. proposed a hybrid framework based on whole-image and patch-based segmentation, based on 3D-UNet and 3D-UNet++ to leverage the benefits of both approaches, while maintaining computational efficiency[17]. An alternative view to the problem of coronary vessel representation comes from graph-based approaches. For instance, Van Herten et al. proposed state-of-the-art unstructured mesh generation methods specifically for patient-specific coronary models[18], while Jia et al. developed a structured mesh generation framework, which supports accurate geometric representation for computational fluid dynamics applications[19]. Hybrid approaches aimed at capitalizing on both the superior boundary precision properties of 2D methods and the contextual integration capabilities of 3D techniques. Beyond the use of ensemble approaches which rely on simple voting mechanisms (e.g., Gan et al.[20]), there is a clear need for systematic integration frameworks, which leverage the complementary advantages of 2D and 3D methods. In light of this, in this research, we propose Seg2RefineNet, a novel two-stage hybrid framework that integrates local spatial information through spatio-frequency attention-based 2D segmentation with global contextual refinement via 3D Attention-GAN processing, achieving both precise vessel boundary extraction and topological consistency preservation.

The core contributions of this research are as follows:

- A spatio-frequency attention-based model is proposed for accurate segmentation of coronary arteries and precise boundary extraction in 2D slices of CCTA images.
- A 3D Attention-GAN-based model is introduced for refinement of coronary artery segmentation, leveraging the principles of image-to-image translation.
- We perform an in-depth analysis of performance contributions made by the architectural components in the proposed framework.
- A thorough comparative performance analysis of Seg2RefineNet w.r.t. state-of-the-art methods, on the largest publicly available dataset, is presented. Moreover, we investigate cases where the proposed model leads to segmentation inaccuracies and identify possible sources of errors.

The remainder of the manuscript is organized as follows. The proposed techniques and strategies are presented in the Methods section. The experimental settings and results are provided in the Experimental Setup and Results sections, respectively. The Discussion section comprehensively analyzes key findings, identifies limitations, and draws conclusions based on experimental results.

Methods

In this study, we propose Seg2RefineNet, a novel framework for coronary artery segmentation using 2D slices of CCTA images followed by 3D volumetric refinement, illustrated in Fig. 1. It consists of two networks, i.e., a spatio-frequency attention-based network (SFANet) for 2D segmentation of coronary arteries and a 3D-Attention GAN-based network for refinement of the 2D coronary artery segmentation. SFANet segments the coronary arteries in a 2D slice-by-slice manner and combines the predicted vessel segmentations in their corresponding 3D volume. Next, the 3D Attention-GAN serves as a volumetric refinement module that processes both the initial segmentation volume, assembled from the slice-wise predictions, and the original CCTA volume to improve segmentation accuracy. The Attention-GAN aims at addressing the inherent limitations of 2D processing by correcting false positives (erroneously segmented non-vessel regions), recovering false negatives (missed vessel segments), enforcing spatial continuity across adjacent slices, and preserving the topological consistency of the coronary arterial tree structure.

In the following subsections, we explain in detail the proposed networks. However, to ensure clarity and consistency throughout the paper, we adopt the following terminology: (i) Seg2RefineNet denotes the overall proposed framework; (ii) SFANet refers to the spatio-frequency attention-based network used for 2D slice-wise segmentation; (iii) 3D Attention-GAN refers to the volumetric refinement module that improves the initial 2D predictions. We denote the input 3D CCTA image as X having spatial dimensions of $H \times W \times D$ with C number of channels. The preprocessed image is represented as X'. The corresponding ground truth label is denoted as T, where each voxel in T holds binary values of 0 and 1, for background (non-vessel) and foreground (vessel), respectively. The initial segmentation map for the CCTA image X is represented as S, while its 3D-refined version is represented as Y.

Spatio-frequency attention-based 2D vessel segmentation

SFANet aims at the task of binary segmentation of coronary arteries in CCTA images at the 2D slice level. For this purpose, it follows a 2D encoder-decoder architecture, as shown in Fig.1 (Top). Given the enhanced 3D CCTA volume X' (obtained as described in data preparation section) with spatial dimensions $H \times W \times D$, we extract

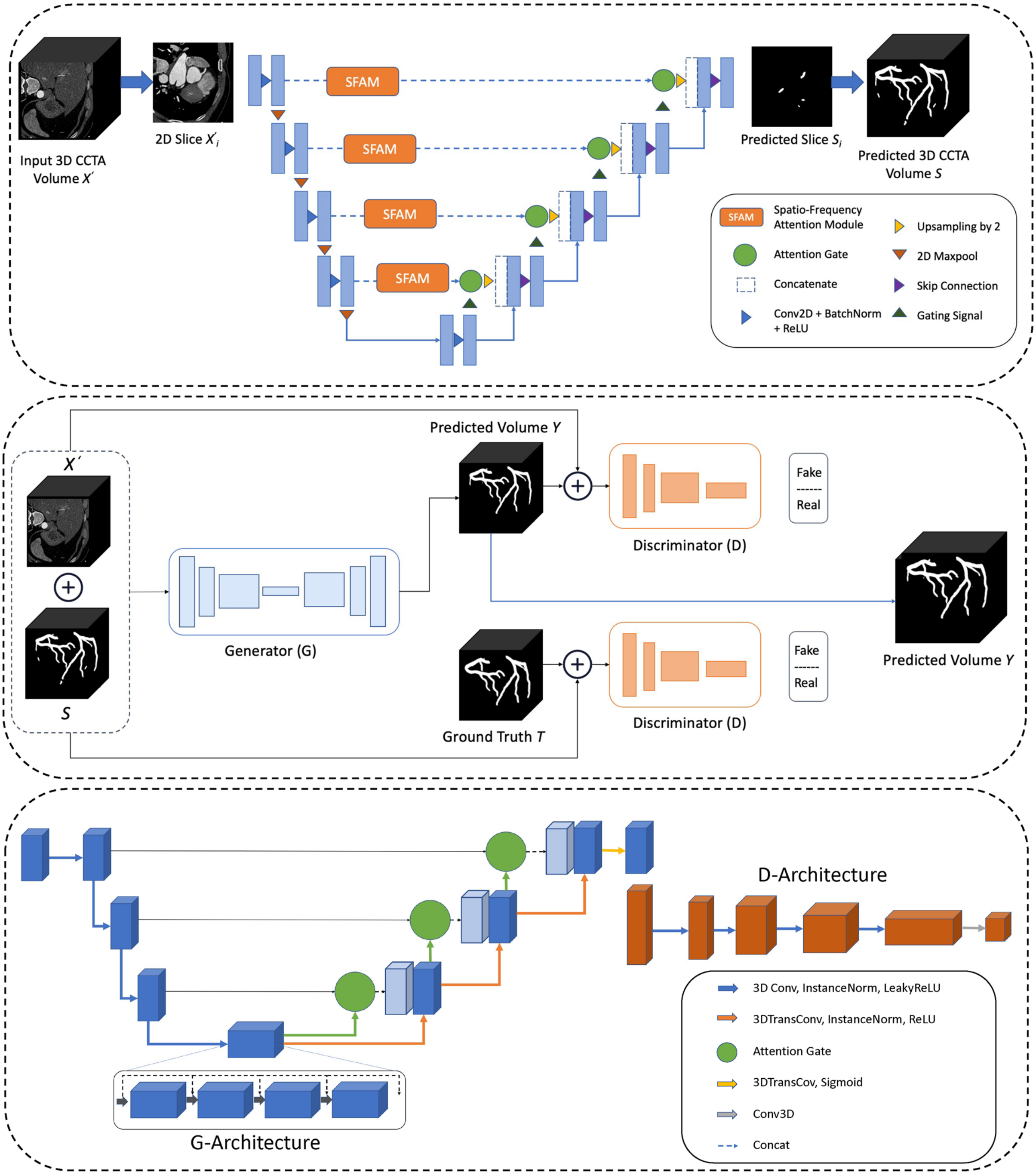

Fig. 1. Illustration of the Seg2RefineNet framework. (Top) Architecture of the spatio-frequency attention-based network (SFANet) for 2D coronary artery segmentation. The input shows the CCTA slices X_i', obtained from the enhanced 3D CCTA volume X' and the output shows the initial 2D segmentation masks S_i, which are concatenated together to produce the initial 3D segmentation volume *S*. (Center) 3D refinement pipeline using Attention-GAN. From left to right: The inputs to the generator are the enhanced CCTA volume X' and the initial segmentation volume *S*, while the discriminator is trained using X', *S*, *Y* and Ground Truth annotations *T*. After GAN processing, the refined 3D segmentation volume *Y* is produced. (Bottom) Detailed generator (G) and discriminator (D) architectures of the 3D Attention-GAN showing downsampling/upsampling blocks, residual connections, and attention mechanisms.

individual 2D slices for processing. Let $X_i' \in R^{(H \times W)}$ represent the i-th axial slice of the enhanced 3D volume X', where $i \in 1, 2,..., D$ denotes the slice index along the depth dimension. The encoder takes each pre-processed 2D CCTA slice X_i' as input and outputs the corresponding binary segmentation mask $S_i \in \{0, 1\}^{(H \times W)}$ from the decoder, where S_i represents the vessel segmentation for the i-th slice. The complete initial 3D segmentation volume *S* is formed by stacking all slice-wise predictions, i.e., $S = \{S_1, S_2, \ldots, S_D\}$.

The encoder is a 5-layered convolution neural network (CNN). Each layer consists of two convolution layers and a max-pooling layer. Features are extracted using a 2D convolution operation (kernel size 3x3), followed by a batch normalization layer (BN) and a nonlinear activation unit, i.e., *ReLU*. Considering C_j as the number of feature channels of the *jth* encoding layer, where $j \in \{1,2...5\}$, $H_j \times W_j$ represents the dimensions of the extracted feature map. The input to the model is a single-channel grayscale image of size $1 \times H_1 \times W_1$, where H_1 and $W_1 = 512$. C_1 to C_5 represent the number of feature channels from the most shallow to the deepest layer, with the number of channels following the sequence from 64, 128, 256, 512, and 1024. We use a maximum pooling layer with a kernel size of 2 x 2 and a step size of 2 for down-sampling the spatial dimensions of the feature map after each subsequent layer.

The decoder focuses on restoring the encoded image features, however, the feature transformation may lead to information loss. This loss may be compensated by fusing the original encoded features into the decoder to restore the target boundaries through skip connections. Coronary arteries, in the 2D slices of a CCTA image, are represented as small foreground segments on a large background of soft tissue, with very little contextual information. As a result, segmentation performance depends on the task-relevant information in the original features. To this end, we take advantage of edges formed due to intensity variations between the coronaries and surrounding soft tissues. For this purpose, we use frequency filtering combined with attention mechanisms to improve the flow of feature information from encoder to decoder, as discussed in the following section.

The decoder network consists of four layers with each layer comprising of two convolutional layers and an upsampling layer. Bilinear interpolation is used in this regard to reduce the number of feature channels and increase the size of the feature map by 2. The upsampled feature map is concatenated with the corresponding features from the encoder. Lastly, a convolution layer (kernel size of 1 x 1) is used to obtain the segmented image of size 1 x H_1 x W_1.

Spatio-frequency Attention Module (SFAM)

Frequency domain analysis has shown significant promise in computer vision tasks. High-frequency components typically correspond to sharp edges and fine details, while low-frequency ones capture global structural information and smooth variations. Recent works demonstrated the advantages of integrating frequency-domain processing with spatial attention mechanisms. Mathai et al. first proposed the use of frequency-based approaches for vessel segmentation in ultrasound imaging[21]. They demonstrated the benefits of preserving high-frequency boundary information, while effectively managing noise artifacts. An interesting development is FcaNet, introduced by Qin et al.[22], which formulates channel attention as a frequency decomposition process using the discrete cosine transform. They proved that conventional global average pooling is a special case of frequency domain feature compression. Rao et al. proposed Global Filter Networks, which learn spatial dependencies in the frequency domain using the discrete Fourier transform and learnable global filters, achieving log-linear computational complexity[23]. Recently, Zhou et al. introduced XNet, a wavelet-based architecture to decompose biomedical images into low and high-frequency components, demonstrating superior segmentation performance through multi-scale frequency feature fusion[24]. Inspired by previous research, the proposed SFAM module utilizes frequency domain decomposition for improved vessel boundary detection in CCTA images. In this context, high-frequency components capture vessel cross-sectional boundaries and intensity variations, while low-frequency components aim at encoding the overall shape of the vessels and contextual information. By focusing the model's attention on task-relevant frequency components and integrating them through channel and spatial attention mechanisms, SFAM achieves improved feature representation for precise coronary artery segmentation, as shown in Fig. 2.

Given the input feature map from the encoder F_e of size $C_j \times H_j \times W_j$, C_j represents the number of channels of *jth* encoding layer and $H_j \times W_j$ are the spatial dimensions of that layer. We apply the 2D discrete Fourier transform independently to each channel along its spatial dimensions. Following concatenation, we denote the resulting feature map in the frequency domain by f_e of size $C_j \times H_j \times W_j$, and proceed by decomposing it into complementary high and low frequency components, f_{high} and f_{low}, respectively. The decomposition is performed channel-wise using two complementary binary masks, M_{high} and M_{low}, of the same size as the spatial dimensions of the feature map (i.e., $H_j \times W_j$). The masks are designed to be mutually exclusive and collectively exhaustive, ensuring that each frequency component is captured by exactly one mask, as follows:

$$M_{high} = 1 - M_{low}, \;\; M_{low} \cup M_{high} = 1, \;\; M_{low} \cap M_{high} = 0. \tag{1}$$

To generate M_{low}, a square of size $(H_j/K) \times (W_j/K)$ is centered in the mask, where K determines the size of the window relative to the spatial dimensions of the feature map. The mask coefficients are assigned a value of 1 within the central region and 0 elsewhere. For instance, when K = 4, M_{low} has a window of size $(H_j/4)$ $\times$ $(W_j/4)$ centered within it, where all values inside the window are set to 1, and all values outside the square are set to 0. M_{low} captures the global structural information, since low frequencies are concentrated toward the center of the frequency-transformed image. Conversely, M_{high} captures the high-frequency components located toward the borders of the 2D frequency spectrum, representing edges and fine details. Channel-wise frequency decomposition is performed as follows:

$$f_{low} = f_e \odot M_{low}, \tag{2}$$

$$f_{high} = f_e \odot M_{high}. \tag{3}$$

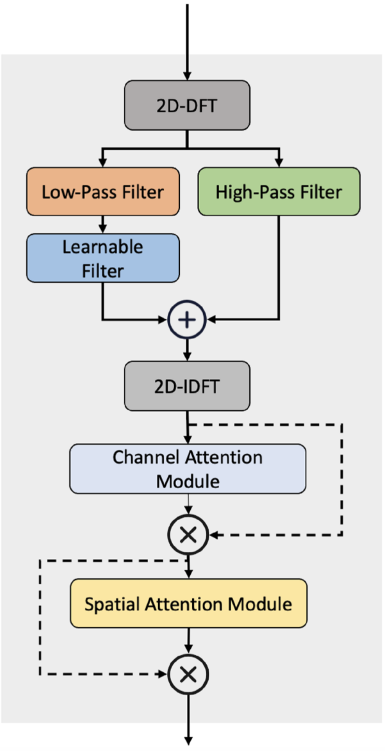

Fig. 2. Block-based representation of Spatio-Frequency Attention Module (SFAM).

A learnable filter L is applied to f_{low} to selectively enhance vessel-relevant frequencies from the low-frequency components. The enhanced frequency representation f' is then constructed by combining the original high-frequency components with the filtered low-frequency components as

$$f' = f_{high} \oplus (f_{low} \odot L)\,, \tag{4}$$

where $\oplus$ and $\odot$ represent element-wise addition and multiplication, respectively. By using the two-dimensional inverse discrete Fourier transform (2D-IDFT), the associated spatial domain representation, F', is obtained. This results in the frequency enhanced representation of the vessel regions within the feature map.

To further enhance feature representation, channel and spatial attention mechanisms are incorporated into the framework[25]. These attention modules allow the model to focus on the most informative regions in the feature map, dynamically emphasizing important channels and spatial locations within those channels.

Spatial information of a feature map is aggregated across channels using both average- and max-pooling operations, generating two spatial context descriptors, i.e., F^c_{avg} and F^c_{max}, denoting the average-pooled and max-pooled feature maps, respectively. Both descriptors are then passed on to a shared network to produce the channel attention map, M_c. The shared network is a multi-layer perceptron (MLP) with one hidden layer. After the shared network processes each descriptor, the output feature vectors are concatenated using element-wise summation. Channel attention is computed, resulting in F'_c as

$$\begin{aligned} F'_c = M_c(F') &= \sigma(MLP(\text{AvgPool}(F')) + MLP(\text{MaxPool}(F'))), \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right), \end{aligned} \tag{5}$$

where W_0 and W_1 are the weights of the MLP shared for both the input descriptors, and σ denotes the sigmoid operation. After the significant channels within the feature map are highlighted, F'_c undergoes spatial attention. Average and max-pooling layers are applied to the input, generating feature maps, $F_{avg} \in R^{1\times H\times W}$ and $F_{max} \in R^{1\times H\times W}$, respectively. These feature maps are processed through a convolution layer to generate the resulting spatial attention map $M_s(F)$ as

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}([\text{Avg Pool}(F); \text{MaxPool}(F)])\right), \\ &= \sigma\left(f^{7\times 7}\left([F_{avg}; F_{\max}]\right)\right), \end{aligned} \tag{6}$$

where $f^{7\times 7}$ denotes the convolution operation with kernel size of 7 × 7.

This refinement of feature information facilitates improved segmentation performance by enabling the network to learn contextually relevant and edge-preserving features, thereby effectively capturing both high-frequency (edge-related) and low-frequency (contextual) components of the image.

Attention-GAN-based volumetric vessel refinement

Generative Adversarial Networks (GANs) demonstrated significant potential in vessel segmentation and refinement tasks. In the context of retinal vessel segmentation, Huang et al. proposed X-GAN, which integrates GANs with vascular biostatistics to achieve near-perfect segmentation accuracy without requiring labeled data[26]. Deng et al. applied cycle Wasserstein GANs with gradient penalty (WGAN-GP) for motion artifact correction in CCTA images, reporting significant improvements in PSNR and clinical quantitative scores[27]. In 3D volumetric vessel segmentation, Sweeney et al. introduced VAN-GAN for 3D vascular network segmentation without the requirement of annotated ground truth data[28]. Gonzales et al. presented a validation study of GAN-based data augmentation approaches for cardiac MRI late gadolinium enhancement segmentation, demonstrating that using GAN-generated synthetic data consistently improves segmentation performance[29]. In a departure from previous GAN-based vessel segmentation approaches that primarily focus on 2D refinement and motion artifact correction, the proposed 3D Attention-GAN specifically addresses the unique challenges of volumetric coronary artery segmentation by incorporating topological consistency constraints and inter-slice relationship modeling. Existing GAN-based methods, e.g., VAN-GAN, directly target 3D vessel segmentation, however, they neither specifically optimize for the refinement of initial 2D segmentations nor incorporate attention mechanisms tailored to coronary vessel characteristics.

The second stage of Seg2RefineNet employs a three-dimensional Attention-GAN designed to refine the initial 3D segmentation map from SFANet. The generator network *G* follows an encoder–bottleneck–decoder structure inspired by the Attention U-Net, extended to volumetric inputs. The input to the generator is the concatenation of the enhanced CCTA volume X' and its corresponding initial segmentation mask *S*, forming a two-channel 3D volume $\tilde{X}$. The encoder of G comprises four successive downsampling blocks. Each block contains a 3D convolution layer with a kernel size of 4 × 4 × 4 and stride 2 × 2 × 2, followed by instance normalization and LeakyReLU activation. To enhance representational capacity and to suppress irrelevant background information, attention gates are integrated between the encoder and decoder stages, allowing skip connections to selectively propagate spatially relevant features. The bottleneck of the generator consists of four residual blocks. Each residual block includes a 3D convolutional layer of kernel size 4 × 4 × 4 and stride 1 × 1 × 1, instance normalization, and LeakyReLU activation, with the block output concatenated with its input to promote gradient flow and stabilize training. The decoder mirrors the encoder with three upsampling stages. Each stage applies a 3D transposed convolution with kernel size 4 × 4 × 4and stride 2 × 2 × 2, followed by instance normalization and ReLU activation. Skip connections from the encoder, modulated by the attention gates, are concatenated with the upsampled feature maps at each stage, ensuring preservation of fine vessel details alongside global context. The final output layer applies a 3D transposed convolution with kernel size 4 × 4 × 4, stride 1 × 1 × 1, and voxel-wise softmax activation, producing refined binary segmentation masks distinguishing vessel and background.

The discriminator *D* is implemented as a 3D PatchGAN, which classifies local volumetric patches as real or fake rather than evaluating the entire volume. Its architecture consists of four downsampling blocks, structurally identical to those in the generator encoder, i.e., 3D convolution with kernel size 4 × 4 × 4, stride 2 × 2 × 2, instance normalization, and LeakyReLU. This design ensures sensitivity to high-frequency structural details critical for vessel boundary accuracy. The final output layer is a 3D convolution with kernel size 4 × 4 × 4 and stride 1 × 1 × 1, followed by sigmoid activation to produce voxel-wise patch-level discrimination scores.

Network training

SFANet is trained using the Adam optimizer with a learning rate of $1e-4$ and a weight decay of $1e-5$. The loss is computed using the sum of the weighted binary cross entropy loss $\mathscr{L}_{wbce}$ and Dice loss $\mathscr{L}_{dice}$. To address the inherent class imbalance in vessel segmentation where background pixels vastly outnumber vessel pixels, we employ $\mathscr{L}_{wbce}$ as

$$\mathscr{L}_{\text{wbce}} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_{pos}\cdot y_i\log\left(p_i\right) + w_{neg}\cdot\left(1-y_i\right)\log\left(1-p_i\right)\right], \tag{7}$$

Where $y_i \in \{0,1\}$ represents the ground truth label for pixel *i*, p_i is the predicted probability, *N* is the total number of pixels, and w_{pos} and w_{neg} are the positive and negative class weights, respectively, calculated based on the vessel-to-background pixel ratio to ensure balanced learning.

Inverse frequency weighting is used with the weights computed as follows:

$$w_{\text{pos}} = \frac{N}{2\times N_{\text{vessel}}}, \tag{8a}$$

$$w_{\text{neg}} = \frac{N}{2\times N_{\text{background}}}, \tag{8b}$$

where N_{vessel} and $N_{background}$ represent the number of vessel and background pixels, respectively.

The SFANet loss function combines the weighted binary cross-entropy loss with Dice loss to handle both class imbalance and boundary precision as follows:

$$\mathscr{L}_{\text{SFANet}} = \alpha\mathscr{L}_{\text{wbce}} + \ (1-\alpha)\mathscr{L}_{\text{dice}}, \tag{9}$$

where α = 0.5 provides equal contribution from both loss terms, as previously employed in[5,30]. The weighted binary cross-entropy $\mathscr{L}_{wbce}$ (Eq. 7) ensures that vessel pixels receive appropriate attention during training despite their scarcity. To improve SFANet's generalization capability and avoid overfitting, geometric 2D data augmentation techniques such as random horizontal and vertical flipping, and random rotation up to 15 degrees are used.

Training of the 3D Attention-GAN involves alternating updates of the *G* and *D* networks. *G* is trained to produce the refined segmented image *Y* by reducing the $\mathscr{L}_{bce}$ as the adversarial loss combined with dice loss $\mathscr{L}_{dice}$ to deceive *D* in distinguishing it from *T*. Differently to the $\mathscr{L}_{wbce}$ in SFANet which assigns higher weight to the vessel pixels and lower to the background, the standard binary cross-entropy (BCE) loss $\mathscr{L}_{bce}$ is applied uniformly across all pixels, when training the 3D Attention-GAN. To ensure that the segmentation results have same topological characteristics as the ground truth, an additional term of the difference in the *Betti Number* (i.e., number of connected components (*CC*)), referred as the topological consistency loss term and denoted by $\Delta\ B_0$ is also added to the generator loss as

$$G_{\text{loss}} = \mathscr{L}_{bce} + \lambda \cdot \mathscr{L}_{dice} + \Delta B_0, \tag{10}$$

$$\Delta B_0 = \frac{\left|B_0^Y - B_0^T\right|}{B_0^Y}, \tag{11}$$

where B_0^Y and B_0^T are, respectively, the Betti numbers of the model prediction *Y* and ground truth segmentation *T*.

$\Delta\ B_0$ serves as a geometric constraint within the loss function, ensuring that the refined segmentation maintains the same number of connected components as the reference annotation, thereby encouraging the generator to preserve the correct arterial tree connectivity. Unlike traditional regularization terms that depend solely on model parameters or predictions to enforce intrinsic properties, $\Delta\ B_0$ is a supervised loss component that explicitly compares topological features against the ground truth. λ is a scalar weight coefficient, originally used in[31], that adjusts the tradeoff between the generator's ability to produce accurate segmentation masks versus the discriminator's ability to guide the generation process.

The loss function of *D* is also made up of two components. The first part is the real discriminator loss (D_{rloss}), which helps *D* to classify real images correctly. The second term is the fake discriminator loss (D_{floss}), which allows *D* to classify the fake/generated images. Both the losses, i.e., D_{rloss} and D_{floss}, are computed by comparing *patchsize* number of true and generated image patches with the corresponding labels (real and fake), respectively. The total discriminator loss (D_{loss}) is then computed as the average of the D_{floss} and D_{rloss} terms as

$$D_{loss} = (D_{rloss} + D_{floss})/2. \tag{12}$$

The generator and discriminator models are trained simultaneously by updating their weights based on their respective losses to improve the overall performance of the 3D-Attention GAN. Algorithmic representation of the training process is presented in the Algorithm 1. Before the training process, data augmentation techniques such as random rotation up to 15 degrees, random masking and addition of gaussian noise were used.

All models were trained on NVIDIA GeForce RTX-4090 GPU.

Experimental setup

This section provides details on the utilized data, evaluation procedure and performance metrics.

Data

To evaluate the performance of the proposed framework, the publicly available ImageCAS dataset was used, comprising of 1000 3D CCTA volumes, each from a unique patient. The data was acquired by a Siemens 128-slice dual-source scanner[17]. All acquisitions were made using high-dose CCTA. The acquired data have sizes of 512 x 512 x (206 - 275) voxels, with a planar resolution of 0.29–0.43 mm^2, and spacing of 0.25–0.45 mm. The data were collected from clinical cases at the Guangdong Provincial People's Hospital during the time period of April 2012 to December 2018. Only patients older than 18 years and with a documented medical history of ischemic stroke, transient ischemic attack and/or peripheral artery disease were eligible for inclusion. For each of the CCTAs, the left and right coronary arteries were independently labeled by two radiologists according to the AHA naming convention[32], and their results were cross-validated. In case of any discrepancy, a third radiologist would perform a further annotation and the final annotation would be determined by consensus.

To assess the generalizability of the proposed framework, we employed the Automated Segmentation of Coronary Arteries (ASOCA) dataset[33], a publicly available benchmark specifically designed for validating coronary artery segmentation algorithms (https://asoca.grand-challenge.org/). ASOCA comprises 40 Cardiac Computed Tomography Angiography (CCTA) scans from 40 unique patients, with a balanced distribution of 20 healthy subjects and 20 patients with confirmed coronary artery disease. This balance ensures robust evaluation across diverse pathological conditions, including both normal vessel morphology and various degrees of coronary stenosis. All CCTA images were acquired using contrast agent administration to enhance vessel-background contrast, following standard clinical protocols for coronary imaging. The dataset includes both proximal and distal coronary segments, encompassing the full spectrum of vessel diameters and anatomical variations encountered in clinical practice. Ground-truth segmentations were produced by three expert annotators working independently, ensuring high-quality and reliable reference standards. The annotation process focused on segmenting the coronary artery lumen, specifically excluding calcified regions, plaque deposits, and other pathological manifestations to maintain consistency with clinical segmentation objectives.

```
Input  : $S, X', T, G, D, patchsize, \lambda, Epochs$
Output : $G$
$real \leftarrow ones(patchsize)$
$fake \leftarrow zeros(patchsize)$
for $i \leftarrow 1$ to $Epochs$ do
    $\tilde{X} = \text{Concat}(S, X')$
    $Y \leftarrow G(\tilde{X})$
    $R \leftarrow D(T, \tilde{X})$
    $D_{rloss} \leftarrow \mathscr{L}_{BCE}(R, real)$
    $F \leftarrow D(Y, \tilde{X})$
    $D_{floss} \leftarrow \mathscr{L}_{BCE}(F, fake)$
    $D_{loss} \leftarrow (D_{rloss} + D_{floss})/2$                        ▷ Eq. (12)
    $D \leftarrow Backprob(D_{loss})$
    $B_0^Y \leftarrow CC(Y)$
    $B_0^T \leftarrow CC(T)$
    $\Delta B_0 \leftarrow (B_0^Y - B_0^T)/B_0^Y$                          ▷ Eq. (11)
    $G_{loss} \leftarrow \mathscr{L}_{bce}(Y, real) + \lambda . \mathscr{L}_{dice}(Y, T) + \Delta B_0$   ▷ Eq. (10)
    $G \leftarrow Backprob(G_{loss})$
end
return $G$
```

Algorithm 1. Training of 3D Attention GAN for Volumetric Refinement

Data preparation

CT images represent the acquired information in Hounsfield Units (HU), which typically range from −1000 to 3000[34]. To visualize specific anatomical structures (in our case, coronary arteries) with their respective densities, a window center of 200 and a window size of 600 is used[10].

To further enhance the cross-sectional representation of the vessel structures in 2D slices, unsharp masking is employed as a pre-processing step. This is an image sharpening technique that employs a Gaussian filter to produce a blurred version of the original image[35]. The smooth version of the image is then subtracted from the original image. The difference is then added to the original image, highlighting edges and high-frequency components. This results in an enhanced version of the original CCTA image, represented as X'.

Evaluation procedure

To train the models in Seg2RefineNet, we used the train-validation-test split, as was proposed in the study that were the first to present and use the ImageCAS dataset[17]. Based on this, the 3D CCTA dataset was split into 700 volumes for training, 50 for validation, and 250 for testing purposes. SFANet was trained on the 2D slices of the corresponding CCTA volumes resulting in the initial 2D segmentations. To avoid any data leakage, during the refinement stage of the initial segmentations, the 3D-Attention GAN was trained on the 50 CCTA volumes of the validation set in the original ImageCAS data split. To ensure a thorough evaluation, this process was repeated using 4-fold cross validation, as originally performed in[17].

Evaluation metrics

To measure the segmentation performance of the proposed models and provide a fair comparison with existing approaches, a common set of metrics is needed. In this regard, based on the metrics used in the state-of-the-art, the dice similarity coefficient (DSC) (or simply, dice score) and Hausdorff distance (HD) are used. The dice score is a measure of similarity between two sets. In the context of image segmentation, it measures the similarity between the predicted segmentation map S or Y and the ground truth T. The mathematical representation of the dice score is given as

$$DSC = \frac{2|Y \cap T|}{|Y| + |T|}. \tag{13}$$

The Hausdorff distance, on the other hand, only considers the pixels at the boundary belonging to the same class. It measures the similarity between the boundaries of the segmented regions compared to those of the ground truth. It calculates the distance between each voxel in the boundaries of model predictions and the corresponding voxels in the ground truth segmentation as follows:

$$\text{HD}\,(Y, T) = \max\left(\max_{t \in T}\min_{y \in Y} d(t, y), \max_{y \in Y}\min_{t \in T} d(y, t)\right), \tag{14}$$

where y and t represent the predicted and ground truth voxels, respectively.

Results

In this section, we provide an analysis of model performance for the SFANet, followed by an ablation study, evaluating the impact of each of the SFANet components on the model output. Next, we present and discuss the performance of Seg2RefineNet, which processes the outputs of SFANet through the 3D-Attention GAN. We proceed with a comparison with competitive 3D and hybrid 2D/3D models in the state-of-the-art, before presenting an analysis of the computational complexity of the proposed model.

2D-based coronary artery segmentation results

Once the slice-by-slice segmentation mask for each input CCTA volume is obtained, it is compared with their corresponding ground truth annotations. Results showed that the SFANet obtained a mean dice score of 0.8024 ± 0.03 as the initial 2D-based segmentation performance. Integration of frequency domain processing in the SFAM module addresses a critical gap in existing 2D approaches. While traditional methods struggle with the uneven contrast distribution characteristics of CCTA images, frequency decomposition enables selective enhancement of vessel-relevant spectral components, yielding improvements in boundary precision, particularly beneficial for detecting small caliber vessels (diameter $< 2\text{mm}$) that are often missed by conventional 2D methods. This choice is further supported in recent research by Alirr et al.[36], which demonstrates that incorporating Hessian-based vesselness preprocessing can improve small vessel detection by up to 1.76%. The proposed frequency domain processing achieves similar benefits through learnable filters that adaptively enhance tubular structures, providing a more flexible alternative to fixed vesselness operators. Another challenge in 2D coronary segmentation is the extreme class imbalance (e.g., vessel pixels $< 3\%$ of total pixel count). The weighted binary cross-entropy formulation specifically addresses this through dynamic weight adjustment based on vessel-to-background ratios. Upon detailed analysis of the performance across different volumes, it was found that SFANet segmented 97.1% of the CCTA volumes with a dice score greater than 0.7. Out of which, 58.8% of the volumes were predicted with a dice score ranging from 0.8 to 0.9. These results show that the SFANet effectively segmented the majority of the volumes in the ImageCAS dataset with a high segmentation performance.

The complexity of the task of segmenting the cross-sectional representations of coronary arteries is not the same throughout the CCTA volume. Results showed that it is dependent on the position of the coronary vessel within the volume. Indeed, variations in vessel thickness and complex branching patterns in the arterial tree, pose challenges in the accuracy of segmentation methods. To this extent, we expanded the analysis of the proposed 2D-segmentation model's performance w.r.t the position of coronary arteries in a CCTA image as

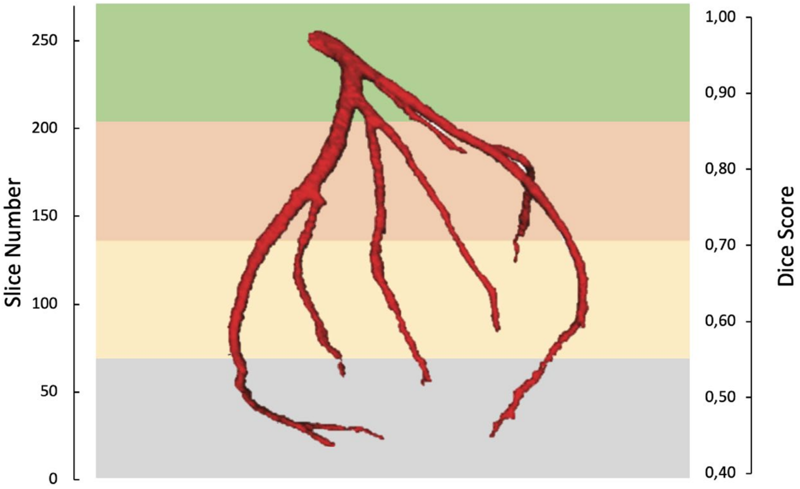

Fig. 3. Average Dice Score achieved using SFANet with respect to the location of the coronary vessel segment within the CCTA volume. Slice numbers represent the axial position from the coronary ostium to the distal vessel terminus (slice 1). The color coding represents Dice score ranges: green (Dice > 0.85, highest accuracy at proximal arterial root), salmon pink (Dice 0.7–0.85, high accuracy at mid segments), yellow (Dice 0.55–0.7, moderate accuracy at distal segments), and grey (Dice < 0.55, lowest accuracy at challenging thin vessel regions). Performance generally decreases from proximal to distal locations due to decreasing vessel diameter and increasing anatomical complexity.

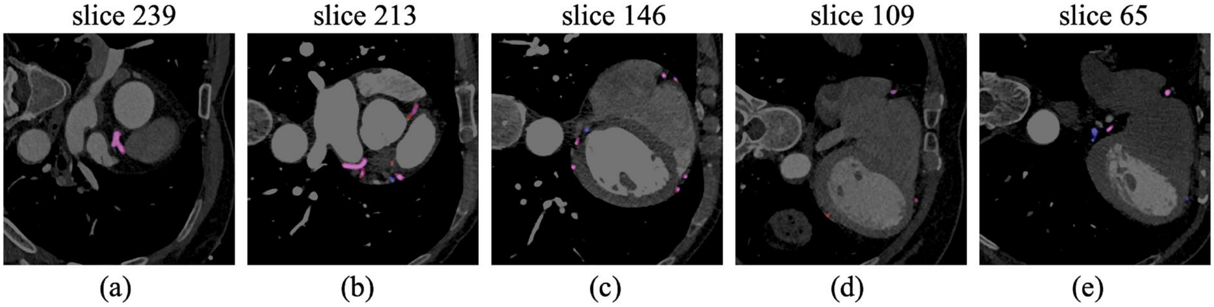

Fig. 4. Segmentation performance w.r.t the coronary artery position represented by slice number. From (**a**-**e**) are slices segmented by SFANet; the colored coded annotations are red for false negatives, blue for false positives, and magenta for correct segmentations.

Model	*Att*	*AG*	*FFE*	$DSC \uparrow$
UNet				0.7643 ± 0.05
AG-UNet	✓			0.7814 ± 0.04
AG-Att-UNet	✓	✓		0.7930 ± 0.04
SFANet	✓	✓	✓	0.8024 ± 0.03

Table 1. Comparison of 2D segmentation models with a vanilla *U*-Net as baseline along with various architectural components on the ImageCAS dataset. *Att* represents the use of both channel and spatial attention mechanisms, *AG* represents the use of Attention Gate, and *FFE* represents frequency feature enhancement. Dice score is used as the metric for comparison purposes.

shown in Fig. 3. To assess segmentation performance relative to anatomical position within the CCTA volume, each axial slice of the 3D CCTA scan is assigned a slice number, starting from the slice containing the distal end of the artery (slice 1) to coronary artery ostium. Average Dice scores are computed per slice across all patient volumes in the test set. This yields a slice-wise average metric indicating segmentation quality at that anatomical location. Slices are grouped into anatomical segments, such as proximal (root), mid, and distal segments, based on known coronary artery branching. The resulting mean Dice scores per slice are then visualized as bars per position to capture spatial performance trends along the arterial tree. This analysis enables identification of challenging anatomical regions where segmentation accuracy decreases due to vessel thinning, branching complexity, or reduced contrast. Results showed that on average, SFANet achieved a high dice score (0.9 to 1.0) when segmenting CT slices closer to the root of the arterial tree. This, on average, amounts to the top 50 to 70 CT slices. Since vessels have a higher diameter and contrast in those slices, a higher segmentation performance is achieved (see Fig. 3). A sample slice is shown in Fig. 4(a) showing the root of the arterial tree as the ground truth. As we can see, with its relatively thick structure and a single point of origin, it is perfectly segmented by the proposed model. On the other hand, as the coronary artery starts to branch out, vessels tend to become thinner and can appear in multiple spatial locations. As a result, the model may struggle to differentiate between the vessel and the surrounding tissue. As a result, a dice score (0.8 to 0.9) is achieved, while segmenting the next 50 to 60 CT slices. In these slices, vessels are still distinctively visible and therefore the majority of the vessel structure is correctly segmented. As can be seen in Fig. 4(b), the model is able to correctly segment the majority of the regions of interest (RoI), with some false positives. Going down a further 50 slices, vessels start to become narrower and appear as small blob-like structures, compared to the surrounding soft tissues. Furthermore, since they appear in multiple spatial locations within each slice, they tend to lose contextual relationships. This results to an increase in the number of false positives and false negatives, which become apparent in the segmentation map. Fig. 4(c-e) show the sparse distribution of the vessel structures with a relatively higher number of false positives and false negatives.

Ablation study

In the previous subsection, we presented the overall 2D-based segmentation performance of the proposed SFANet. Here, we present the ablation study, which illustrates the impact to the model performance of the various architectural components, when added to a baseline U-Net architecture, resulting in the construction of SFANet.

As shown in Table 1 (row 1), we started with the 5-layered vanilla U-Net as a baseline architecture to test model performance. A mean dice score of 0.7643 ± 0.05 was achieved, indicating a promising segmentation performance, leaving room for potential improvement. Sample segmentation maps of 2D-slices are shown in the Fig.5(b) and (g). Compared to the vanilla U-Net, attention gate U-Net (AG-UNet) introduces a single attention mechanism in the form of attention gate at the skip connection between encoder to decoder. This allowed the selective screening of class-relevant information, further improving the dice score to 0.7814 ± 0.04, see

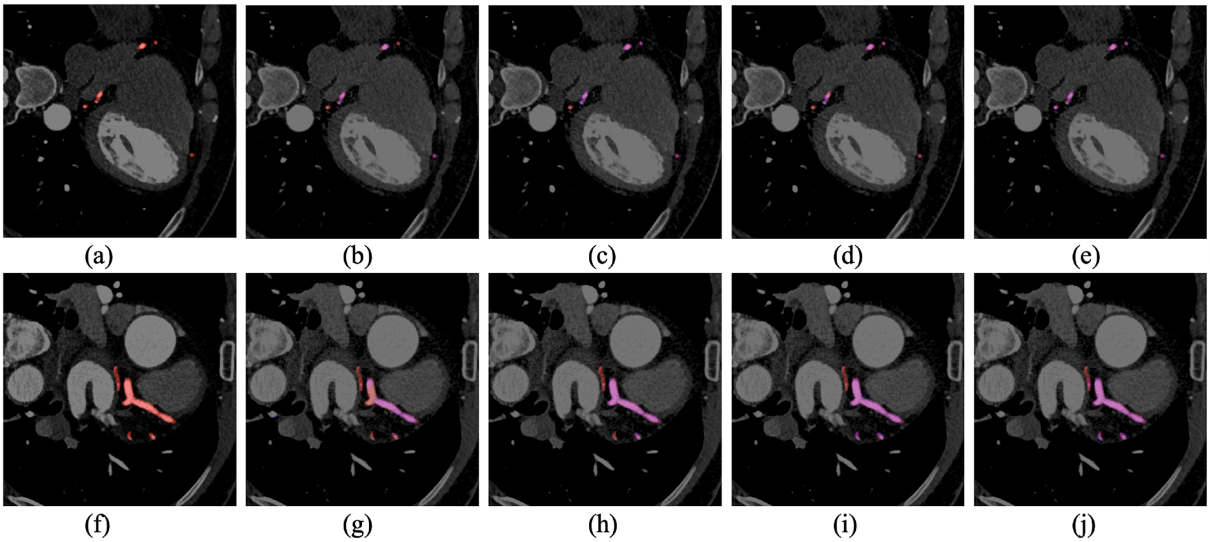

Fig. 5. Segmentation performance comparison for the ablation experiments in two 2D CCTA slices. From left to right are ground truth, UNet, AG-UNet, AG-Att-UNet, and SFANet (proposed model), respectively; The colored annotations are orange for ground truth, blue for model prediction, and magenta for the overlap of ground truth and prediction.

Table. 1 (row 2). Fig.5(c) and (h) show visual improvements in the segmentation performance, where AG-UNet accurately segmented the vessel structure, which was left as false negative by the UNet. To further strengthen feature selection, channel and spatial attention are progressively incorporated, resulting in AG-Att-UNet. By applying channel attention, the model is able to select channels with significant information, limiting the impact of less informative channels. Afterwards, a layer of spatial attention is added to exploit the spatial relationships within channels, thus allowing the model to localize important regions within the feature maps. The attention gates then selectively allow relevant features from channel and spatially attentive feature maps to pass through to the decoder, enhancing the representation of important areas and suppressing irrelevant information. This addition led to a further increase in the segmentation performance with a mean dice of 0.7930 ± 0.04, see Table.1 (row 3). Fig. 1(d) and (i) show an increase in the model's capability to capture almost all of the vessel structures.

Although attention mechanisms effectively filtered and enhanced the quality of encoded features, a further enhancement to the feature maps as input to the attention blocks was proposed in this research. Specifically, spatial frequencies may represent various anatomical features. By enhancing relevant frequency components, it is possible to effectively provide useful information before it is processed by the channel and spatial attention blocks. The application of the attention blocks to the frequency-enhanced feature maps leads to SFANet, resulting to the highest mean dice score of 0.8024 ± 0.03, compared to the previous architectures (see Fig.5(e) and (i)).

3D-based coronary artery segmentation results

With the input CCTA X', the initial segmentation masks from SFANet resulted to a mean dice score of 0.8024. Sample 3D representations of the initial segmentation (S), are shown in Fig.6(b) and (f). Compared to the ground truth (T) shown in the Fig.6(a) and (e), we can see that the initial segmentation was able to capture the overall topological structure of coronary arteries. However, as we saw in the 2D segmentation results, false positives in the slices capturing the branching of the arterial tree are profoundly represented in 3D. This is likely due to the lack of contextual information across different slices, particularly, in the later part of the volume. Alongside the false positives, parts of the vessel structure were not included in the initial 2D-based segmentation, representing false negatives. Therefore, the initial 2D segmentation results show potential room for improvement towards a finer segmentation of coronary arteries.

To provide the 3D-Attention GAN a more guided approach towards refining the initial segmentation mask, alongside the 2D-based vessel segmentation volume, the original CCTA volume is provided. This allows the model to capture contextual information across the entire CCTA volume, thus, learning to reduce the occurence of false positives and predict the missing false negatives, due to topological inconsistency. Indeed, the 3D-Attention GAN improved the initial 2D-based segmentation results from a mean dice score of 0.8024 to 0.8313. Sample evidence for this improvement is shown in Fig.6(c) and (f), where the model not only successfully removed false positives but also retrieved missing information in the initial segmentation.

Comparative analysis with the state-of-the-art

To draw a fair comparison with the state-of-the-art methods, in this sub-section, we considered those evaluated on the ImageCAS dataset. In this regard, we include both single step-based methods such as 3D-FCN[9] and 3D-UNet[15], and multi-step methods of 3D-UNet and 3D-UNet++[17], ensemble of 2D and 3D-UNets[20], and CFNet[37] as baseline models.

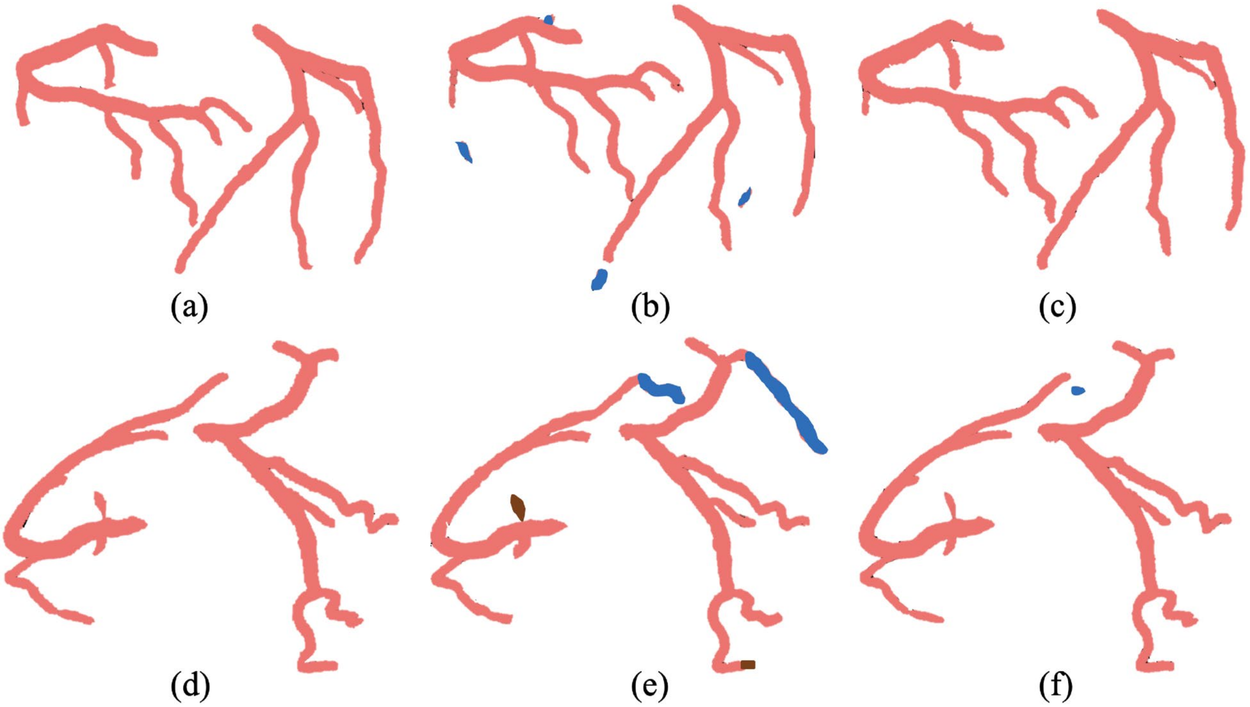

Fig. 6. Sample results for 3D Segmentation. (**a**) and (**d**) show the ground truth annotations (T), (**b**) and (**e**) represent the initial segmentation (S) while the (**c**) and (**f**) show the refined segmentation mask (Y). The colored annotations describe the false negatives in brown, and false positives in blue.

The classification results of the baseline models reported on the ImageCAS dataset are shown in Table 3, rows 1–10. 3D-FCN[9] achieved a mean dice score of 0.8058 with HD of 28.66 mm. These results are achieved by directly processing 3D CCTA volumes. When processed using a 3D-UNet based architecture[15], a lower performance is reported with a dice score of 0.7201 and HD of 40.96 mm.

Aiming to boost segmentation performance, one of the multi-step approaches combines the coarse segmentation results from the 3D-UNet model and fine patch-based segmentation results using 3D-UNet++[17]. Next, a variety of patch sizes were used to train the 3D-UNet++, with the results combined in an ensemble setting. This approach achieved a mean dice score of 0.8296 with HD of 27.21 mm. CFNet followed a similar approach by performing coarse segmentation using a 3D-UNet, followed with fine segmentation using a 3D transformer-based architecture[37]. The model achieved a mean dice score of 0.8267 with HD of 18.83 mm. To leverage local and global contextual information from 2D- and 3D-based methods, an ensemble based approach was proposed, which combines predictions following a voting-based approach[20]. Although the method did not outperform[17] in terms of dice score, it achieved a lower HD, indicating improved segmentation of the vessel boundaries. DiffCAS[38], a diffusion-based multi-attention network, reported a DSC score of 84.59% with HD of 11.92 mm on ImageCAS dataset. Similarly, SADiff[39] combined spatial attention with a diffusion generator and, achieved a mean dice score of 83.48% and HD of 19.43 mm. AGFA-Net[11] on the other hand reported the highest dice score on the ImageCAS dataset (i.e., DSC = 86.74%) together with a Hausdorff distance of 0.23 mm. Although these methods[11,38,39] reported high performance, it is important to note that the evaluation was carried out using random data splits rather than the standardized splits published in the original article[17]. Finally, for completeness, a recent 3D-PSPNet variant trained on a 200-case ImageCAS subset reported a DSC of 0.76 (using global processing), which illustrates the range of recent results when non-official splits or smaller subsets are used. Although the performance of Seg2RefineNet can be seen as lower than the DiffCAS[38] and AGFANet[11] and close to the SADiff[39], its results are reported following the official data split[17]. Therefore, overall, Seg2RefineNet remains competitive on ImageCAS achieving an average dice score of 0.8313 ± 0.018 and HD of 12.95 ± 0.53 mm.

Considering the best performance achieved on official data split, among the baseline approaches, in terms of the dice score and HD, the proposed Seg2RefineNet framework outperformed existing state-of-the-art methods with a mean dice score of 0.8313, and mean HD of 12.95 mm.

Computational complexity

To evaluate if the performance gain of Seg2RefineNet compared to the existing methods is achieved at the cost of increasing computational complexity, the computational complexity of the proposed model is analyzed. Table 2 provides an overview of computational complexity of Seg2RefineNet and the existing methods.

3D-FCN[9], 3D-UNet[15], 3D-UNet & 3D-UNet++[17], 3D-PSPNet[40] are purely 3D convolutional approaches, thus exhibiting $O(H \times W \times D \times K^3)$ complexity, with the cubic kernel term dominating the computational

Sr.No	Method	Input Data	DSC ↑	HD (mm) ↓	Complexity
1.	3D-FCN	3D	0.8058	28.66	$O(H\times W\times D\times K^3)$
2.	3D-UNet	3D	0.7201	40.96	$O(H\times W\times D\times K^3)$
3.	3D-UNet & 3D-UNet++	3D	0.8296	27.21	$O(H\times W\times D\times K^3)$
4.	2D+3D-UNet Ensemble	2D & 3D	0.8231	17.54	$O(H\times W\times D\times K^3)$
5.	CFNet	3D	0.8267	18.83	$O(N^2)$
6.	DiffCAS	3D	0.8459	11.92	$O(I\times H\times W\times D\times K^3)$
7.	SADiff	3D	0.8348	19.43	$O(I\times H\times W\times D\times K^3)$
8.	AGFANet	3D	0.8674	0.23	$O(N^2)$
9.	3D-PSPNet	3D	0.76	–	$O(H\times W\times D\times K^3)$
10.	**Seg2RefineNet**	**2D & 3D**	$\mathbf{0.8313 \pm 0.018}$	$\mathbf{12.95 \pm 0.53}$	$O(H\times W\times D\times K^3)$

Table 2. Comparative analysis of segmentation methods evaluated on the ImageCAS dataset.

Sr.No	Method	Training Dataset	Testing Dataset	DSC ↑	HD (mm) ↓	Generalization Gap DSC/HD
1.	3D-UNet	ImageCAS	ASOCA	0.651	45.23	0.069/4.27
2.	CFNet	ImageCAS	ASOCA	0.704	38.67	0.126/19.84
3.	**Seg2RefineNet**	ImageCAS	ASOCA	$\mathbf{0.767 \pm 0.0.056}$	$\mathbf{29.19 \pm 0.71}$	$\mathbf{0.064}$/16.24

Table 3. Cross-Dataset Generalizability Performance.

cost. These techniques process entire volumes through 3D convolutions at each layer. Methods combining 2D and 3D processing, i.e., 2D+3D-UNet Ensemble[20] and the proposed Seg2RefineNet, show $O(H\times W\times D\times K^2)$ $+O(H\times W\times D\times K^3)$ complexity. However, the 3D term $O(H\times W\times D\times K^3)$ dominates over the 2D term $O(H\times W\times D\times K^2)$, thus maintaining cubic complexity but with reduced constants. Specifically, Seg2RefineNet initially processes the 2D slices of the input CCTA volume with time complexity of $O(H\times W\times D\times K^2)$, and then combined with the 3D Attention GAN, it is therefore computationally efficient compared to purely 3D convolutional methods. Although the method in[20] also utilized both the 2D and 3D UNet-based models, training of multiple models in an ensemble learning setup increases the overall computational complexity. CFNet turns out to be the most computationally expensive method. In the case of transformer and attention-based methods, CFNet[37] and AGFA-Net[11], $O(N^2)$ complexity is added to $O(H\times W\times D\times K^3)$ to account for self-attention mechanisms, where N represents the number of patches. In the case of 3D volumes, $N=(H/p)\times(W/p)\times(D/p)$, where p is the patch size. Diffusion models, DiffCAS[38] and SADiff[39], multiply the base complexity by the diffusion iteration count I, resulting in $O(I\times H\times W\times D\times K^2)$. Typical diffusion processes require I = 50–1000 steps, significantly increasing computational cost. In summary, through its hybrid 2D-3D design, Seg2RefineNet demonstrates important computational efficiency, achieving state-of-the-art segmentation performance, while maintaining the same time complexity class as state-of-the-art methods, making it suitable for real-time clinical applications.

Generalizability analysis

To further assess how well Seg2RefineNet performs on out-of-distribution data, we evaluated our model on the Automated Segmentation of Coronary Arteries (ASOCA) Challenge dataset[33]. Seg2RefineNet was trained on ImageCAS and was then tested on the ASOCA dataset in coronary artery segmentation, producing a mean dice score of 0.767 ± 0.056 and an HD of 29.19 ± 0.71 mm. Table 3 systematically compares cross-dataset generalization performance across state-of-the-art approaches trained on the ImageCAS dataset and applied for zero shot performance evaluation on the ASOCA challenge dataset.

Seg2RefineNet demonstrates promising zero-shot generalization with only 6.4% DSC performance drop, with cross-dataset robustness exceeding the remaining two ImageCAS-trained methods by 0.5–6.2% DSC. In regards to HD, 3D-Unet demonstrates the smallest generalization gap with a drop of 4.27 mm compared to ImageCAS, however, its accuracy in both dataset is quite low. CFNet on the other hand has a HD drop of 19.84 mm, demonstrating significant performance deterioration on the ASOCA dataset. Fig. 7 shows sample segmentations (e-h) of the ASOCA dataset compared to clinical annotations (a-d). It can be observed that similarly to the case of the ImageCAS dataset, Seg2RefineNet was able to capture the overall structure of the coronary arteries however, it did struggle in the case of fine thin vessels in the ASOCA dataset, particularly as we moved further away from the root of the arterial tree. This resulted in under-segmentation of the arterial tree.

Compared to the recent studies that utilized ASOCA dataset for training, Qiu et al.[12] reported a Dice score of 0.8853 and an HD of 1.07 mm. Similarly, Yan et al.[41] achieved a Dice score of 0.837 and an HD of 3.72 mm. In contrast, our model when only tested on the ASOCA dataset, without being trained on it, demonstrated competitive segmentation performance. This highlights its strong generalizability when tested on unseen dataset.

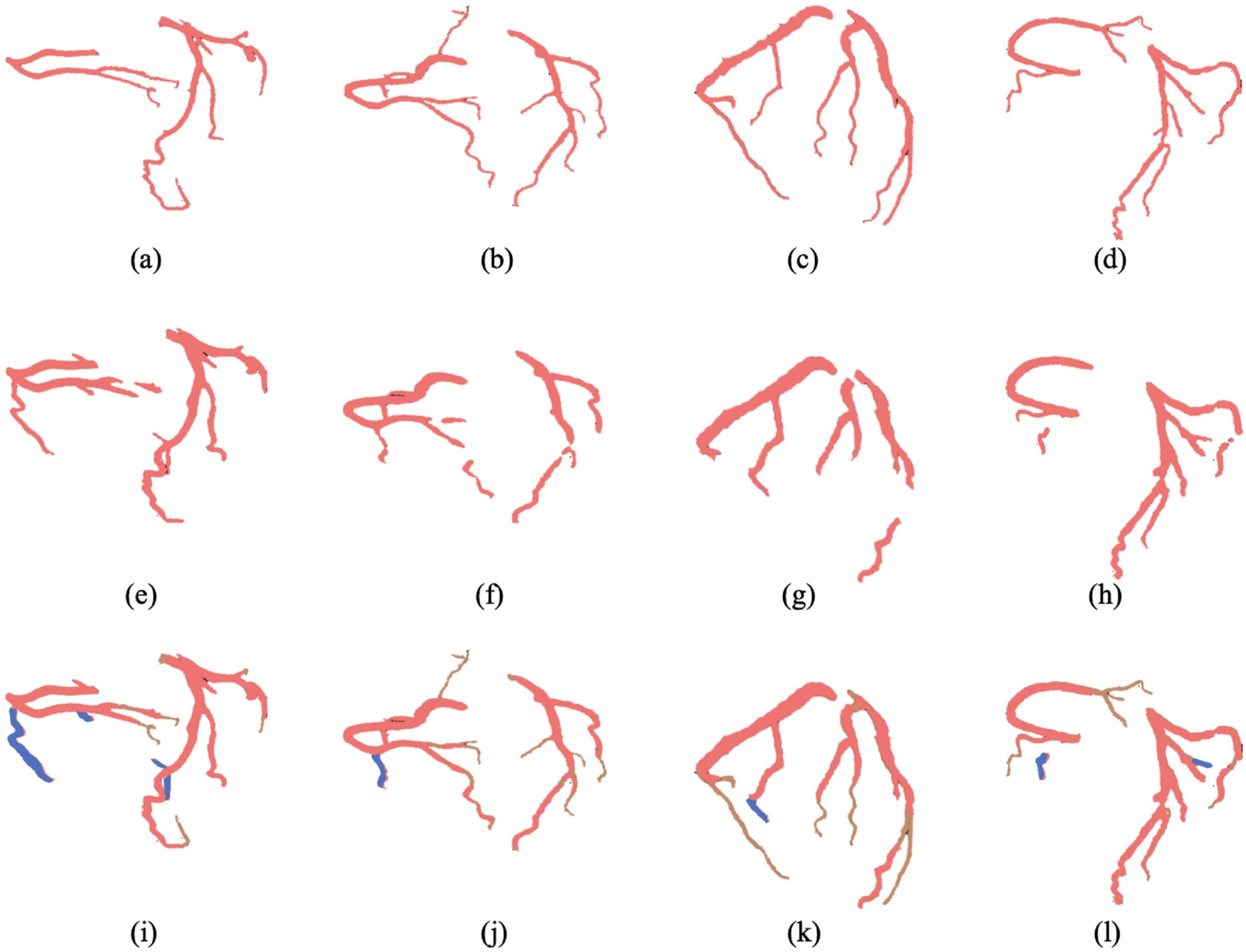

Fig. 7. Sample test results for 3D Segmentation on ASOCA dataset. (**a**-**d**) show the ground truth annotations of ASOCA dataset, (**e**-**h**) represent the segmentation results of our proposed Seg2RefineNet. (**i**-**l**) represent the difference between the ground truth and the segmentation results. False positives are highlighted in blue whereas the false negatives in brown.

Discussion

To shed light into model's performance, we conducted a systematic analysis of CCTA instances where Seg2RefineNet produces lower than expected DSC scores (between 0.6 to 0.7). A few samples of those results with their corresponding ground truth is shown in Fig.8. The discussion of these results has been organized with respect to vessel morphology, vessel intensity and anatomical location. It was observed that the geometric characteristics of the arterial tree play a significant role on segmentation performance. Specifically, the thinning of distal vessels, complex branching topology and the degree of vessel tortuosity. Smaller distal vessels, often characterized by diameters close to CT's spatial resolution limit, can form disconnected components, thus affecting the model's performance. Furthermore, segmentation errors were observed in the vicinity of junction points, i.e., bifurcations and trifurcations, particularly evident when moving from proximal to distal segments, where the model exhibited geometric discontinuities in some of the predictions (see Fig.8(e)). Lastly, while Seg2RefineNet performs very well in straighter segments, it may produce discontinuous segmentations in highly tortuous vessels, and can often result in false positives (see Fig.8(f)). Varying vessel intensity can also contribute to imperfect segmentations. It appears that the model is finding it challenging to deal with gradually diminishing intensity gradients in distal segments resulting to under-segmentations. Despite the use of the spatio-frequency mechanism, blurry vessel boundaries also contribute to segmentation errors. Another pattern, which our analysis revealed, is that lower contrast between the vessels and the background, i.e., lower signal-to-noise (SNR) ratio, or equivalently, higher noise levels contribute to segmentation inaccuracies. This effect is witnessed in both Fig.8(e) and (f).

Conclusion

In this study, we proposed Seg2RefineNet, a two stage coronary artery segmentation method that is developed to achieve accurate segmentation with precise segmentation of vessel boundaries. To this extent, we proposed SFANet, a novel 2D spatio-frequency attention-based UNet architecture equipped with frequency feature enhancement followed by the channel and spatial attention mechanism. Frequency feature enhancement

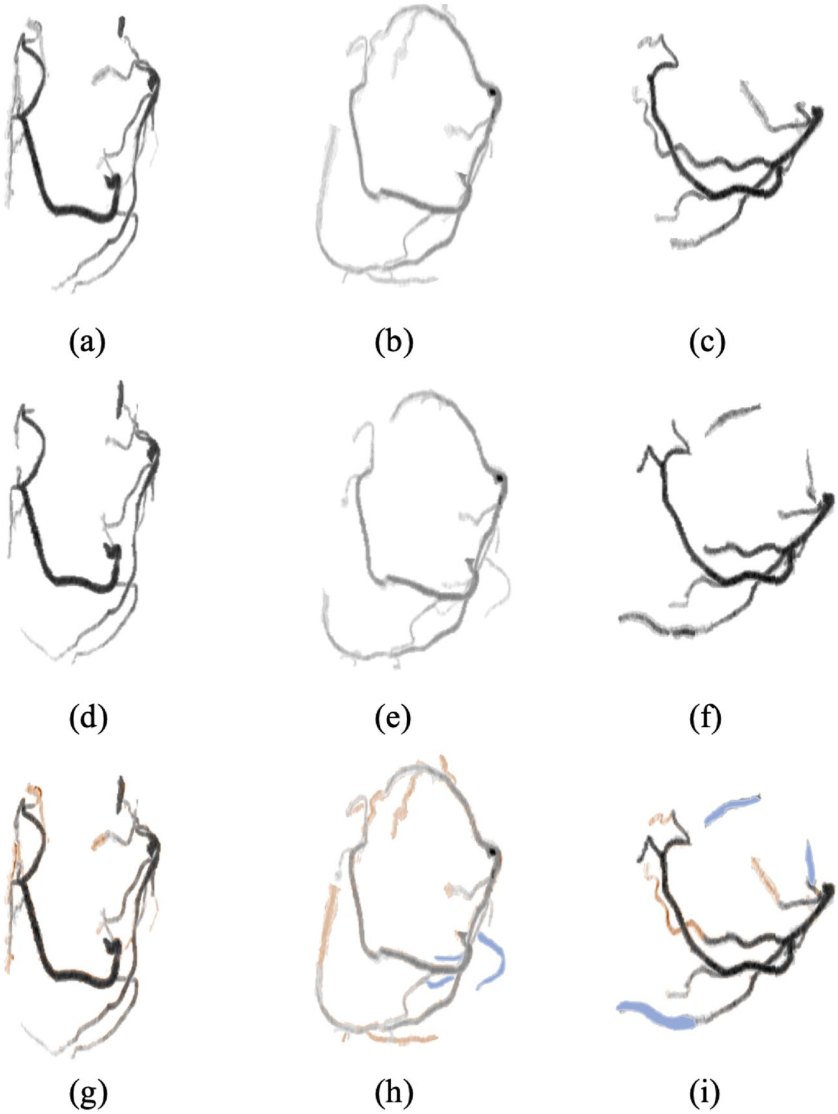

Fig. 8. Representative challenging cases from ImageCAS dataset showing Maximum Intensity Projection (MIP) visualizations with inverted intensity values for enhanced contrast. Ground truth vessels (**a**-**c**) and corresponding segmented results (**d**-**f**) demonstrate cases with Dice scores between 0.6–0.7. The apparent elongated vessel morphology results from MIP projection effects and complex 3D-to-2D visualization constraints. Shades of black highlight the thicker vessels in darker shade while the thinner ones in lighter. (**g**-**i**) represent the differences between the ground truth and predicted vessels. False positives are highlighted in blue and false negatives in brown.

worked by selecting vessel-related information in both high and low-frequency components. Filtered frequency components then undergo the channel and spatial attention allowing the model to focus on the target information. Results showed that although attention itself improved the segmentation performance, compared to the standard UNet, when combined with the frequency feature enhancement, a mean Dice score of 0.8024 was achieved. When it comes to volume-based performance, 97.1% of the CCTA volumes were predicted with a dice score greater than 0.7. Furthermore, compared to the existing methods, which utilized a single step approach, SFANet outperformed the 3D-UNet (dice score 0.7201) and achieved comparable performance to that of 3D-FCN (dice score 0.8058). It was also found that although the method was able to capture the cross-sectional representation of coronary arteries in 2D-slices, it did struggle with handling instances of vessel branching and thin vessels, thus losing the contextual information. This led to the use of a 3D-Attention GAN-based method as a second stage of Seg2RefineNet to refine the initial segmentation results. To ensure the topological consistency in the refined segmentation masks, the difference in the number of connected components compared to the ground truth annotations was also integrated as a loss term during the training of the model. The use of 3D-Attention GAN not only integrated the contextual information across the entire volume but also its generative capabilities allowed the model to learn to refine the vessel segmentation by removing the falsely predicted and generating the falsely removed vessel structures. In Seg2RefineNet, the use of frequency feature enhancement integrated with attention mechanisms allowed the model to focus on the vessel structures, thus leading to precise segmentation of vessel boundaries along with the topological consistency introduced by the 3D-Attention GAN. This combination allowed the model to achieve a mean dice score of 0.8313 with HD of 12.95 mm. This not only ensured the accurate segmentation of both vessel structures and their boundaries, but also outperformed the existing state-of-the-art methods. When tested on ASOCA dataset, a mean dice score of 0.767 and HD of 29.19 mm, showed that Seg2RefineNet is able to generalize well on unseen data.

In the future, we will consider the integration of methods that can capture small vessels effectively resulting in a better segmentation performance with even lower HD. To this extent, we aim to incorporate vessel-oriented filters that can help to highlight vessels and capture the cross-sectional information of the coronary artery more effectively. Work can also be done to incorporate a multi-level attention mechanism that can leverage the information from the deep and shallow layers at the same time. Integrating transformer-based architectures as generators within the GAN framework can be explored as part of our future work to further improve long-range dependency modeling and global context understanding. This could be particularly beneficial in complex 3D segmentation tasks. We also aim to design a custom task-specific loss function that is sensitive to the vessel boundaries and variations and to thoroughly evaluate and fine-tune its integration into our method. Lastly, we would like to evaluate the Seg2RefineNet on external/real world datasets to further assess its generalization capability.

Data availability

The dataset used in this study is a publicly available dataset ImageCAS (https://www.kaggle.com/datasets/xiaoweixumedicalai/imagecas)

Code availability

Source code used in this work is available for non-commercial purposes from the corresponding author on request.

Received: 5 May 2025; Accepted: 16 October 2025
Published online: 20 November 2025

References

1. Di Cesare, M. et al. World heart report 2023: Confronting the world's number one killer. *World Hear. Fed.* (Geneva, Switzerland, 2023).
2. Shao, C., Wang, J., Tian, J. & Tang, Y.-d. Coronary artery disease: from mechanism to clinical practice. *Coron. Artery Dis. Ther. Drug Discov.* 1–36 (2020).
3. Jia, D. & Zhuang, X. Learning-based algorithms for vessel tracking: A review. *Comput. Med. Imaging Graph.* **89**, 101840 (2021).
4. Cheung, W. K. et al. A computationally efficient approach to segmentation of the aorta and coronary arteries using deep learning. *Ieee Access* **9**, 108873–108888 (2021).
5. Hong, P. et al. A u-shaped network based on multi-level feature and dual-attention coordination mechanism for coronary artery segmentation of ccta images. *Cardiovasc. Eng. Technol.* **14**, 380–392 (2023).
6. Chen, F. et al. Positive-unlabeled learning for coronary artery segmentation in ccta images. *Biomed. Signal Process. Control.* **87**, 105473 (2024).
7. Fu, X. et al. 3dgr-car: Coronary artery reconstruction from ultra-sparse 2d x-ray views with a 3d gaussians representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 14–24 (Springer, 2024).
8. Lei, Y. et al. Automated coronary artery segmentation in coronary computed tomography angiography (ccta) using deep learning neural networks. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11318, 279–284 (SPIE, 2020).
9. Shen, Y. et al. Coronary arteries segmentation based on 3d fcn with attention gate and level set function. *Ieee Access* **7**, 42826–42835 (2019).
10. Wang, Q. et al. Automatic coronary artery segmentation of ccta images using unet with a local contextual transformer. *Front. Physiol.* **14**, 1138257 (2023).
11. Liu, X. & Zhao, C. Agfa-net: Attention-guided and feature-aggregated network for coronary artery segmentation using computed tomography angiography. arXiv preprint arXiv:2406.08724 (2024).
12. Qiu, Y. et al. A topology-preserving three-stage framework for fully-connected coronary artery extraction. *Med. Image* 103578 (2025).
13. Zhang, X. et al. An anatomy-and topology-preserving framework for coronary artery segmentation. *IEEE Transactions on Med. Imaging* **43**, 723–733 (2023).
14. Kong, B. et al. Learning tree-structured representation for 3d coronary artery segmentation. *Comput. Med. Imaging Graph.* **80**, (2020).
15. Chen, Y.-C. et al. Coronary artery segmentation in cardiac ct angiography using 3d multi-channel u-net. arXiv preprint arXiv:1907.12246 (2019).
16. Huang, W. et al. Coronary artery segmentation by deep learning neural networks on computed tomographic coronary angiographic images. In *2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 608–611 (IEEE, 2018).
17. Zeng, A. et al. Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Comput. Med. Imaging Graph.* **109**, 102287 (2023).
18. van Herten, R. L. et al. World of forms: Deformable geometric templates for one-shot surface meshing in coronary ct angiography. *Med. Image Analysis* 103582 (2025).
19. Jia, D. et al. Dvasmesh: Deep structured mesh reconstruction from vascular images for dynamics modeling of vessels. In *International Workshop on Graphs in Biomedical Image Analysis*, 118–128 (Springer, 2024).
20. Gan, M., Xie, W., Tan, X. & Wang, W. Coronary artery segmentation framework based on three types of u-net and voting ensembles. *Heal. Inf. Sci. Syst.* **13**, 1–10 (2025).
21. Mathai, T. S., Jin, L., Gorantla, V. & Galeotti, J. Fast vessel segmentation and tracking in ultra high-frequency ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 746–754 (Springer, 2018).
22. Qin, Z., Zhang, P., Wu, F. & Li, X. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 783–792 (2021).

23. Rao, Y., Zhao, W., Zhu, Z., Lu, J. & Zhou, J. Global filter networks for image classification. *Adv. neural information processing systems* **34**, 980–993 (2021).
24. Zhou, Y., Huang, J., Wang, C., Song, L. & Yang, G. Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21085–21096 (2023).
25. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).
26. Huang, C. et al. X-gan: A generative ai-powered unsupervised model for main vessel segmentation of glaucoma screening. arXiv preprint arXiv:2503.06743 (2025).
27. Deng, F. et al. Image restoration of motion artifacts in cardiac arteries and vessels based on a generative adversarial network. *Quant. Imaging Medicine Surg.* **12**, 2755 (2022).
28. Sweeney, P. W. et al. Unsupervised segmentation of 3d microvascular photoacoustic images using deep generative learning. *Adv. Sci.* **11**, 2402195 (2024).
29. Gonzales, R. A. et al. Quality control-driven deep ensemble for accountable automated segmentation of cardiac magnetic resonance lge and vne images. *Front. Cardiovasc. Medicine* **10**, 1213290 (2023).
30. Yang, G. & Pan, B. Skin lesion image segmentation algorithm based on mc-unet. *IEEE Access* (2025).
31. Cirillo, M. D., Abramian, D. & Eklund, A. Vox2vox: 3d-gan for brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, 274–284 (Springer, 2021).
32. Cerqueira, M. et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. *Circulation* **105**, 539–542 (2002).
33. Gharleghi, R. et al. Automated segmentation of normal and diseased coronary arteries-the asoca challenge. *Comput. Med. Imaging Graph.* **97**, 102049 (2022).
34. Razi, T., Emamverdizadeh, P., Nilavar, N. & Razi, S. Comparison of the hounsfield unit in ct scan with the gray level in cone-beam ct. *J. dental research, dental clinics, dental prospects* **13**, 177 (2019).
35. Vincent, D. J., et al. Edge enhancement and noise smoothening of ct images with anisotropic diffusion filter and unsharp masking. In *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 55–59 (IEEE, 2018).
36. Alirr, O. I., Al-Absi, H. R., Ashtaiwi, A. & Khalifa, T. Efficient extraction of coronary artery vessels from computed tomography angiography images using resunet and vesselness. *Bioengineering* **11**, 759 (2024).
37. He, S. et al. Cfnet: A coarse-to-fine framework for coronary artery segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 431–442 (Springer, 2023).
38. Li, J. et al. Diffcas: diffusion based multi-attention network for segmentation of 3d coronary artery from ct angiography. *Signal, Image and Video Processing* **18**, 7487–7498 (2024).
39. Xu, R., Dai, L., Wang, J., Zhang, L. & Wang, Y. Sadiff: Coronary artery segmentation in ct angiography using spatial attention and diffusion model. *J. Imaging* **11**, 192 (2025).
40. Chachadi, K., Nirmala, S. & Netrakar, P. G. Automated coronary artery segmentation with 3d pspnet using global processing and patch based methods on ccta images. *Cardiovasc. Eng. Technol.* 1–15 (2025).
41. Yang, J. et al. Hwa-resmamba: automatic segmentation of coronary arteries based on residual mamba with high-order wavelet-enhanced convolution and attention feature aggregation. *Phys. Medicine & Biol.* **70**, 075013 (2025).

Acknowledgements

The study is funded by Khalifa University under the grant number RIG-2023-050.

Author contributions

P.L. conceived the study, generated the hypotheses, and provided supervision support. U.K. performed all the experiments, analyzed the data, and wrote the manuscript. P.L. and U.K. contributed to the revision of the manuscript.

Funding

The study is funded by Khalifa University under the grant number RIG-2023-050.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.