Medical Image Analysis 97 (2024) 103265

Contents lists available at ScienceDirect

Medical Image Analysis

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

DCCAT: Dual-Coordinate Cross-Attention Transformer for thrombus segmentation on coronary OCT

Miao Chu [a,b,c], Giovanni Luigi De Maria [b,c,d], Ruobing Dai [a], Stefano Benenati [b,c,e], Wei Yu [a], Jiaxin Zhong [f,a], Rafail Kotronias [b,c,d], Jason Walsh [b,c,d], Stefano Andreaggi [b,g], Vittorio Zuccarelli [b], Jason Chai [b,c], Oxford Acute Myocardial Infarction (OxAMI) Study investigators [b,c], Keith Channon [b,c,d], Adrian Banning [b,c,d], Shengxian Tu [a,c,*]

[a] Biomedical Instrument Institute, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] Oxford Heart Centre, Oxford University Hospitals NHS Trust, UK
[c] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, UK
[d] National Institute for Health Research, Oxford Biomedical Research Centre, UK
[e] University of Genoa, Genoa, Italy
[f] Department of Cardiology, Fujian Medical University Union Hospital, Fujian, China
[g] Division of Cardiology, Department of Medicine, University of Verona, Italy

ARTICLE INFO

Keywords:
Acute coronary syndromes
Optical coherence tomography
Thrombus segmentation
Cross-attention

ABSTRACT

Acute coronary syndromes (ACS) are one of the leading causes of mortality worldwide, with atherosclerotic plaque rupture and subsequent thrombus formation as the main underlying substrate. Thrombus burden evaluation is important for tailoring treatment therapy and predicting prognosis. Coronary optical coherence tomography (OCT) enables in-vivo visualization of thrombus that cannot otherwise be achieved by other image modalities. However, automatic quantification of thrombus on OCT has not been implemented. The main challenges are due to the variation in location, size and irregularities of thrombus in addition to the small data set. In this paper, we propose a novel dual-coordinate cross-attention transformer network, termed DCCAT, to overcome the above challenges and achieve the first automatic segmentation of thrombus on OCT. Imaging features from both Cartesian and polar coordinates are encoded and fused based on long-range correspondence via multi-head cross-attention mechanism. The dual-coordinate cross-attention block is hierarchically stacked amid convolutional layers at multiple levels, allowing comprehensive feature enhancement. The model was developed based on 5,649 OCT frames from 339 patients and tested using independent external OCT data from 548 frames of 52 patients. DCCAT achieved Dice similarity score (DSC) of 0.706 in segmenting thrombus, which is significantly higher than the CNN-based (0.656) and Transformer-based (0.584) models. We prove that the additional input of polar image not only leverages discriminative features from another coordinate but also improves model robustness for geometrical transformation.Experiment results show that DCCAT achieves competitive performance with only 10% of the total data, highlighting its data efficiency. The proposed dual-coordinate cross-attention design can be easily integrated into other developed Transformer models to boost performance.

1. Introduction

Acute coronary syndromes (ACS) lead to life-threatening medical emergencies and remain one of the leading causes of death worldwide. Contemporary imaging studies are shedding new light on the underlying mechanisms of ACS. Atherosclerotic plaque rupture with in situ thrombus formation accounts for the majority of ACS (Souteyrand et al., 2015). Meanwhile, thrombus triggered by plaque erosions are attributed to around 30% of ACS cases (Jia et al., 2013) and may be on the rise in an era of intensive lipid-lowering therapy. ACS could also occur without evidence of thrombus formation. The distinction of different mechanisms behind ACS by assessing the magnitude of thrombotic response influences treatment therapy (Jia et al., 2017). Therefore, accurate quantification of thrombus is of clinical importance in supporting individual optimal therapy according to the principles of personalized and precision medicine (Crea and Libby, 2017).

Intracoronary thrombus burden is semi-quantitatively estimated based on angiography using methods like the Thrombolysis In

* Correspondence to: Room123, Med-X Research Institute, Shanghai Jiao Tong University, No. 1954, Hua Shan Road, Shanghai 200030, China.
E-mail addresses: giovanniluigi.demaria@ouh.nhs.uk (G.L. De Maria), sxtu@sjtu.edu.cn (S. Tu).

https://doi.org/10.1016/j.media.2024.103265
Received 4 January 2024; Received in revised form 2 June 2024; Accepted 1 July 2024
Available online 5 July 2024

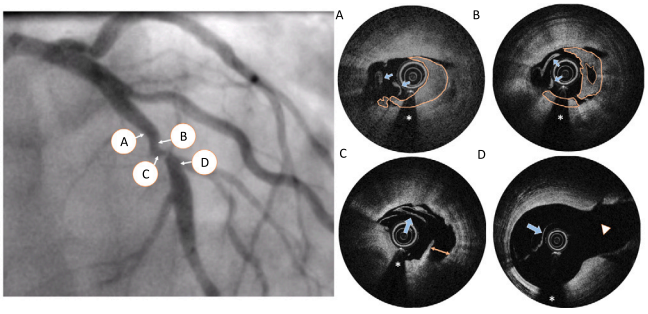

Fig. 1. Example of an ACS case with thrombus formation observed in angiography (left) and OCT (right). Contrast filling deficiency on angiography indicates thrombus, which is clearly visualized on OCT with ruptured intimal (orange arrow in C) and thrombus clots (orange contours in A and B) attached to the lumen. Guide wire shadow and side branch are indicated with star and arrowhead, respectively. Insufficient blood flush during imaging results in artifacts (blue arrow).

Myocardial Infarction (TIMI) thrombus grading and Thrombus Score (TS) (Aleong et al., 2009). However, these methods are subjective. A quantitative approach, named dual QCA (Vergallo et al., 2019), has been developed to refine assessments, but it still cannot provide an accurate evaluation because angiography quantifies thrombus indirectly based on the vessel's contrast-filling appearance. Intravascular imaging of intravascular ultrasound (IVUS) and optical coherence tomography (OCT) have emerged as complementary diagnostic tools to overcome angiography shortcomings by enabling intraluminal visualization of the arteries. The choice between them depends on the specific clinical scenario (Giacoppo et al., 2024): IVUS offers better penetration that allows plaque burden assessment; On the other hand, OCT provides superior spatial resolution (10–20 μm), which is 10× higher than IVUS (≈150 μm). Thus OCT enables in-vivo visualization of thrombus that cannot otherwise be achieved by other image modalities. As shown in Fig. 1, image features of ruptured plaque with thrombus are distinctive on OCT, which are however subtle and hard to quantify on conventional angiography. Several methods have been proposed to quantify thrombus on OCT (Porto et al., 2015). For instance, Prati et al. developed a score based on the quadrants encroached by thrombus within an OCT pullback. However, current methods are limited to manual analysis (De Maria et al., 2017). Given the large frame numbers of one OCT assessment (300–500 frames/pullback), manual analysis poses a great challenge for clinical routine where rapid decision-making is necessary. Hence, automatic segmentation of thrombus on OCT is urgently needed.

The major challenge of segmenting thrombus lies in the complicated features of OCT in ACS patients, where plaque rupture causes discontinuity of intact lumen surface, and thrombus show a large variation in size and shape with extremely irregular boundaries (Fig. 1). Additionally, a common imaging artifact of residual blood may resemble thrombus, where global context is needed for discrimination. Therefore, both global context and local details are essential in the task. Furthermore, although the clinical utilization of OCT is increasing as it is favored by mounting evidence from recent randomized clinical trial (Holm et al., 2023), the current data on OCT is relatively small compared with other medical imaging modalities. Data-efficient deep learning (DL) network is necessary for achieving robust performance.

A unique property of coronary OCT is that the image is initially acquired by intracoronary transducer in a manner of helical scanning, generating raw data in polar coordinate. The images are then transformed into Cartesian coordinate to be in line with the anatomical structure of the vessel. OCT images in Cartesian and polar coordinates offer advantages in identifying various objects, owing to distinct features present in each coordinate system. Previous studies choose one of the two coordinates and completely discard the other. Li et al. (2021) propose novel data augmentation on polar coordinate to generate realistic OCT images, which are converted back to Cartesian coordinate for model training. The potential benefits of complementary features offered by both coordinates have yet to be thoroughly explored.

To address the above challenges, we propose a novel dual -coordinate cross-attention transformer network, referred to as DCCAT. Our main contributions are: (1) We implement automatic thrombus segmentation on coronary OCT for the first time by leveraging cross-attention mechanisms to facilitate feature communication between Cartesian and polar coordinates. In this way, discriminative features in each view can be used to enhance the other via long-range dependence modeling, resulting in refined semantic understanding; (2) We address the data efficiency challenge of Transformer and robustness to geometric transformations by using polar transformation and shift equivalence of convolutional operations. (3) The proposed dual-coordinate cross-attention design can be flexibly extended to concurrent DL models to improve performance. (4) The superiority, robustness and data efficiency of the designed model have been extensively evaluated against state-of-the-art methods using an independent external testing data set.

2. Related works

Since our segmentation framework is mainly built on a cross-attention scheme within OCT images from two coordinates, we herein provide detailed reviews of the related DL works in coronary OCT image analysis, multi-view learning and geometric transformation.

2.1. Coronary OCT analysis

Previous DL studies on coronary OCT mainly focus on classifying different phenotypes of atherosclerotic plaque. Gessert et al. (2018a) applies multiple convolutional neural network (CNN) architectures, including Resnet50, Resnet101, InceptionV3 and Inception-ResnetV2, for frame-wise plaque classification. In a later study (Gessert et al., 2018b), they conduct a comprehensive exploration of transferring learning and data augmentation, which show Cartesian OCT images lead to a superior performance than polar images and benefit substantially more from data augmentation. However, the complementarity offered by the two coordinates is not exploited. Semantic segmentation at pixel level enables detailed plaque characterization. In the early stage, machine learning methods based on handcrafted features are the common pipeline (Athanasiou et al., 2014; Huang et al., 2018; Ughi et al., 2013). Later studies propose different DL methods for plaque characterization. A two-stage DL method is designed for segmentating calcified plaque (Li et al., 2021). The study combines 2D and 3D convolution for in-plane and through-plane feature extraction in the segmentation stage, which is subsequently followed by a classification model to remove false positives. As opposed to the two-stage workflow, our previous study (Chu et al., 2021) proposes an end-to-end deep CNN model, which is trained by a hybrid loss for multi-class segmentation. Pseudo-3D input of stacked consecutive cross-sections is fed and forwarded through the encoding path at multi-scale to integrate spatial information from adjacent images and avoid the computational complexity of 3D convolution. Recently, Park et al. (2022) develop a Transformer-based DL model for the diagnosis of plaque erosion on OCT. Information from adjacent images is integrated by the Transformer model, which outperforms single frame-based CNN model. It is one of the few studies applying Transformer on coronary OCT. However, it is a classification task and self-attention layers are added only on the highest layer of CNN features.

2.2. Cross-attention for multi-view images DL

Previous studies have proven combining divergent information from multi views could improve the semantic understanding of target objects

(Liu et al., 2022; Yan et al., 2020; Yang et al., 2020; Xu et al., 2022). In general, multi-view DL refers to feeding models with data captured from different modalities or sources. The Cartesian and Polar OCT in the study can be considered as two views of image by projecting one coordinate to another. Therefore, we focus on DL models with multi-view images.

The key to multi-view images DL lies in the effectiveness of feature fusion, which is challenging due to potential non-alignment within features from different views. CNN-based model typically adopts one convolutional network for each view separately and then fuses the extracted features together for the remaining part of the network (Feng et al., 2018; Sun et al., 2020; Wei et al., 2019). The fusion scheme involves concatenation, summation, or max pooling, etc.

With the successful application of Transformer in computational vision, feature fusion has been implemented by cross-attention mechanism. Transformer leverages self-attention mechanism to weigh the importance of features according to the relevance of each feature to others. The first vision Transformer (ViT) is proposed by Dosovitskiy et al. by dividing images into patches, where self-attention is applied across flatten patches to explore feature relations. Multi-view images DL has adopted the core-concept of self-attention into cross-attention to mine long-range dependences among features from different views. In a study (Chen et al., 2021a), the input image is partitioned into patch tokens at two sizes, which are then fused by cross-attention to complement each other. The main contribution of this work is to develop multi-scale feature fusions for vision Transformer, yet it is specially designed for classification task. The extension of the work to segmentation needs further investigation.

Cross-attention mechanism overcomes the lack of non-local correlation by CNNs. However, current multi-view images DL usually utilizes cross-attention at the deepest latent features (Lu et al., 2020; Lei et al., 2020; Ahn et al., 2023). For instance, cross-attention is used to improve the segmentation coherence among two consecutive slices by Lei et al. (2020). The workflow results in one input being predicted twice independently, the average of which is used for the final output. Similarly, Ahn et al. (2023) uses a cross-attention transformer network at the end of two encoding branches to extract inter-frame dependent features, which are subsequently used for motion tracking. Similar mechanism has also been applied to extract correlation among video frames (Lu et al., 2020). In contrast to the above approaches where cross-attention is only added at the deepest layer of the encoder, our work applies cross-attention in a hierarchical way for comprehensive feature communication at multiple scales.

2.3. Geometric transformation

Extracting robust features that remain invariant or equivariant to geometric transformation is crucial for DL models' semantics understanding. The success of CNN in visual tasks is partially attributed to the translational equivariance property provided by the convolution operation. However, CNN does not have invariance or equivariance to other geometric transformations, such as rotation and scaling (Jaderberg et al., 2015; Cohen and Welling, 2016; Chen et al., 2018a). Therefore, specialized convolutional or pooling operations have been introduced to address it (Jaderberg et al., 2015; Zhou et al., 2017; Sifre and Mallat, 2013). For instance, deformable convolution network enables the model to learn offsets based on the input to adapt the receptive field to irregular shapes (Dai et al., 2017). Another approach is employing coordinate transformation as a pre-processing step. In Chen et al. (2018a), two polar transformations are proposed to improve rotate-invariant image representations. The combination of coordinate transformation of the input image and regular CNN can be considered as a warped convolution without introducing additional parameters or complexity to the network structure (Henriques and Vedaldi, 2017; Remmelzwaal et al., 2020).

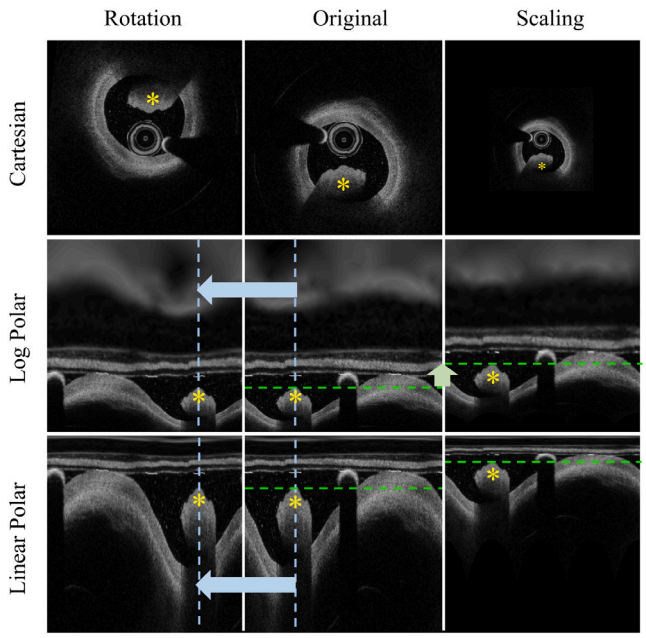

Fig. 2. OCT image transformation between Cartesian and polar coordinates. Rows from top to bottom stand for images in Cartesian, log-polar and linear-polar coordinate, respectively. The middle column is the original image, with the rotation and scaling of the image on the left and right side. Thrombus is indicated by yellow stars. Blue and green arrows indicate horizontal and vertical shifting, respectively.

On the other hand, Transformers lack inductive bias, leading to a high reliance on large scale training data (Azad et al., 2023). Studies have shown that training with extensive data can trump inductive bias, enabling Transformer to outperform CNN (Kolesnikov et al., 2020). However, the data hunger nature of Transformer poses challenges in medical images, where data acquirement is expensive. To address this, hybrid structures have been proposed to combine the advantages of CNN and Transformer, such as UNETR (Hatamizadeh et al., 2022) and TransUNet (Chen et al., 2021b). Nevertheless, we argue that incorporating geometric equivalence into Transformer could further alleviate high data demand.

3. Methods

3.1. Coordinate transformation

We adopt two coordinate transformations, i.e., log-polar transformation and linear-polar transformation, for performance comparison. Images in 2D Cartesian coordinate (x, y) are mapped to log-polar coordinate (ρ, θ) according to Eq. (1)

$$\begin{aligned} \rho &= log(\sqrt{(x-x_c)^2+(y-y_c)^2}) \\ \theta &= atan2(\frac{y-y_c}{x-x_c}) \end{aligned} \tag{1}$$

(x_c, y_c) is the center coordinate of the Cartesian image. The reverse transformation from log-polar to Cartesian coordinate is

$$\begin{aligned} x &= e^{\rho} cos(\theta) + x_c \\ y &= e^{\rho} sin(\theta) + y_c \end{aligned} \tag{2}$$

Applying a scaling of $\Delta\rho$ and rotation of $\Delta\theta$ to the point (x, y) in the Cartesian coordinate results in a new point (x', y'):

$$\begin{aligned} x' &= e^{\rho+\Delta\rho} cos(\theta+\Delta\theta) + x_c \\ y' &= e^{\rho+\Delta\rho} sin(\theta+\Delta\theta) + y_c \end{aligned} \tag{3}$$

As shown in Fig. 2, the central Cartesian image pixels are sampled more aggressively than those at the ends, thus the objects within the

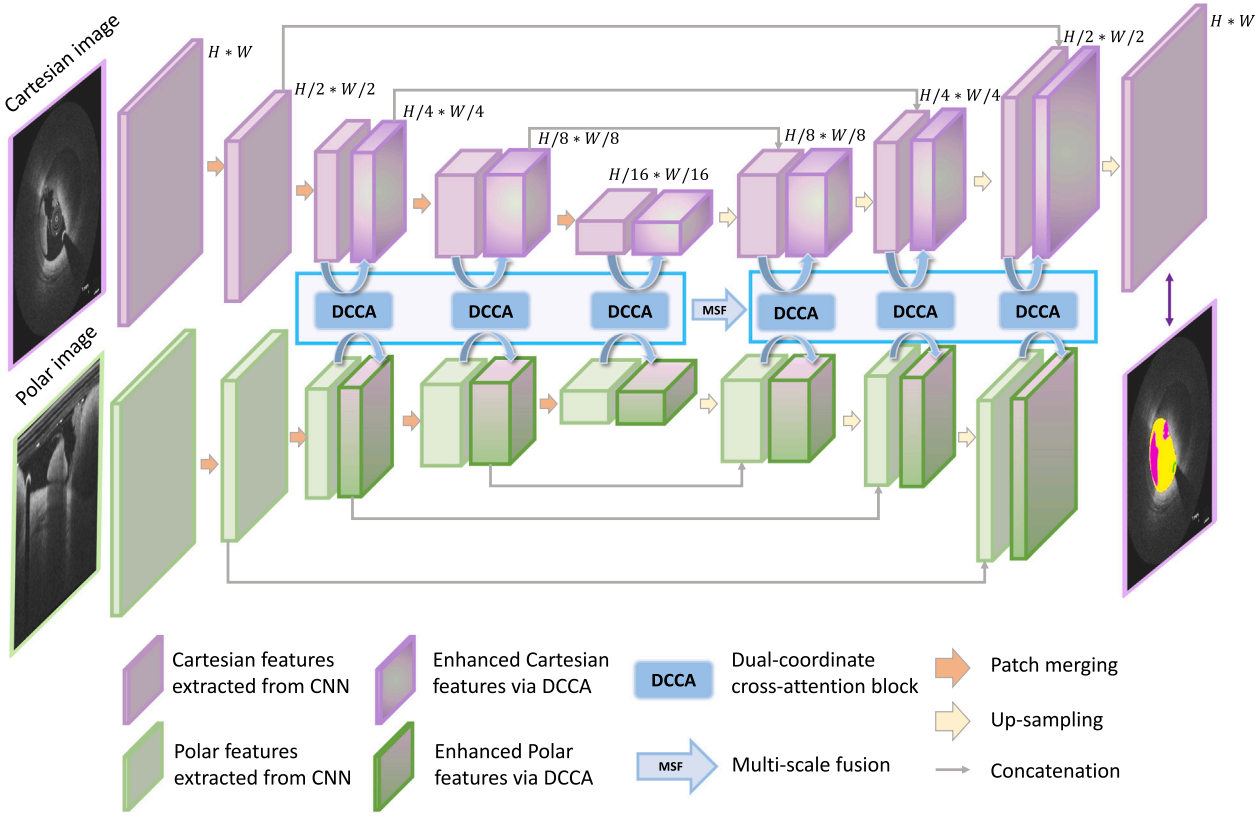

Fig. 3. Overview of the proposed dual-coordinate cross-attention Transformer model. Cartesian and polar images are processed separately by a mirrored two-stream U-shape convolutional network, with DCCA block stacked hierarchically at multi-scale for long-range dependency modeling across the two streams.

vessel wall are compressed to a relatively small portion of the whole image. Therefore, we also adopt linear-polar transformation according to Eq. (4), for the purpose of performance comparison.

$$\rho = \sqrt{(x - x_c)^2 + (y - y_c)^2}$$
$$\theta = atan2(\frac{y - y_c}{x - x_c}) \tag{4}$$

3.2. Proposed framework

3.2.1. Overall structure

Fig. 3 depicts the schematic diagram of the proposed model. The overall architecture follows a U-shape design by adapting it to a mirrored two-stream structure. Specifically, the input Cartesian and polar OCT images are processed by separate convolutional blocks to extract local features. The basic convolutional block is a repeated structure of 'Conv-BatchNorm(BN)-ReLu' with a residual connection:

$$F^{l+1} = F^l + ReLU(BN(conv(F^l))) \tag{5}$$

F^l refers to features at l layer. The feature communication between the two streams is implemented by the proposed dual-coordinate cross-attention (DCCA) block, where long-range dependencies of the two streams are extracted to enhance each other. After fusion, the enhanced features leveraged from the two coordinates are passed to their own convolutional layers to further extract higher-level local features. DCCA is hierarchically stacked within multiple scales throughout the model. Moreover, a multi-scale fusion (MSF) block is utilized to further fuse features from DCCA blocks in the encoding phase, the outputs of which are skip-connected to the corresponding decoder phase. Finally, a convolutional head is applied to Cartesian features at the full resolution, followed by 1×1 convolution layer and SoftMax activation function, generating a multi-category probability map. The model generates only Cartesian output to ensure output consistency and reduce computational complexity. In the next section, we elaborate on the detailed design of the DCCA block.

3.2.2. DCCA block

As shown in Fig. 4, the DCCA block comprises multi-head cross-attention (MHCA), self-attention (SA) and convolutional-based feed-forward network (con-FFW). MHCA utilizes cross-attention mechanisms to fuse features from Cartesian and polar coordinates, aiming to enhance relevant features from both inputs and boost semantic understanding. SA uses self-attention mechanisms within polar features to generate concise tokens, aiming to reduce the computational complexity of MHCA to a linear level.The two-stream features F_c and F_p extracted from Cartesian and polar coordinates are communicated and leveraged according to:

$$\widetilde{F_p} = SA(F_p)$$
$$F_c', F_p' = MHCA(Norm(F_c), Norm(\widetilde{F_p}))$$
$$F_c'' = con\text{-}FFW(F_c' + F_c) \tag{6}$$
$$F_p'' = F_p' + \widetilde{F_p}$$

$Norm$ denotes normalization, F_c'' and F_p'' are enhanced output features. $con\text{-}FFW$ replaces linear layers in the feed-forward network with convolutional layers. In the following paragraphs, we illustrate the design of SA and MHCA in detail.

Inspired by efficient multi-head attention from MedFormer (Gao et al., 2022), self-attention (SA) module is adopted to generate concise token maps for polar features $\widetilde{F_p} = SA(F_p)$. The detailed design of the SA module is shown in Fig. 5. Specifically, a weight map $W \in R^{hw \times H \times W}$ is generated from $F_p \in R^{C \times H \times W}$ by applying convolution and SoftMax layers, where C, H and W represent the channel number,

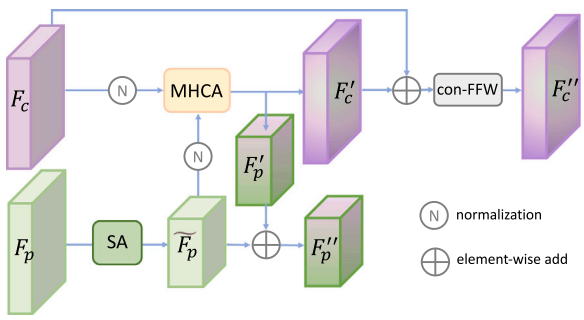

Fig. 4. Schematic illustration of the proposed dual-coordinate cross-attention (DCCA) block. MHCA: multi-head cross-attention; SA: self-attention; conv-FFW: convolutional-based feed-forward network.

height and width of feature map that input into SA, and h and w are the reduced height and width of the output concise feature map.

$$W = SoftMax(Conv(F_p)) \tag{7}$$

Then F_p and W are flattened into shape of $F_p \in R^{C\times HW}$ and $W \in R^{hw\times HW}$, respectively, which are used to generate the concise token map $\widetilde{F_p} \in R^{C\times h\times w}$ by a dot production and reshape.

$$\widetilde{F_p} = Reshape(Flatten(F_p) \cdot Flatten(W)^T) \tag{8}$$

By assigning a fixed size of h and w, where $hw \ll HW$, the concise token map reduces the computational complexity of the subsequent cross-attention to a linear level.

In the MHCA module (Fig. 6), convolutions are firstly applied to project Cartesian F_c and concise polar features $\widetilde{F_p}$ to Query, Key and Values embeddings. The dot product of Query and Key from the two coordinates extracts the dependencies between Cartesian and polar features, which remain consistent regardless of the pair used. To further improve computational efficiency, only one pair (Q_c, K_p) along with each corresponding Value, are used to exchange information between each other, i.e., $Q_c, V_c \in R^{d\times H\times W}$ and $K_p, V_p \in R^{d\times h\times w}$, where d represents the embedding dimension, H and W are the height and width of Cartesian feature map, and h and w are the height and width of the concise polar feature map. The embeddings are then flattened and reshaped into sequences of $R^{HW\times d}$ and $R^{hw\times d}$, and are split into k parallel heads Q_{c_i}, V_{c_i} and K_{p_i}, V_{p_i} $(i = 1, 2, \dots, k)$, with a dimension of $R^{HW\times d_k}$ and $R^{hw\times d_k}$, where $d_k \times k = d$.

In each cross-attention head, weight matrix $CA_i \in R^{HW\times hw}$ is calculated by a scaled dot-production of Q_{c_i}, K_{p_i} to measure the similarity between Cartesian and polar features.

$$CA_i = SoftMax(\frac{Q_{c_i} K_{p_i}^T}{\sqrt{d_k}}) \tag{9}$$

The weight matrix CA_i is subsequently used to weigh V_{c_i} and V_{p_i} to aggregate context information.

$$\begin{aligned} F'_{c_i} &= CA_i \cdot V_{p_i} \\ F'_{p_i} &= CA_i^T \cdot V_{c_i} \end{aligned} \tag{10}$$

where $F'_{c_i} \in R^{HW\times d_k}$ and $F'_{p_i} \in R^{hw\times d_k}$. Furthermore, the outputs of all cross-attention heads are concatenated and once again projected by 1×1 convolutional layer, resulting in the final output of F'_c and F'_p with shape of $R^{d\times H\times W}$ and $R^{d\times h\times w}$.

$$\begin{aligned} F'_c &= Conv(Reshape(Concat(F'_{c_1}, F'_{c_2}, \dots, F'_{c_k}))) \\ F'_p &= Conv(Reshape(Concat(F'_{p_1}, F'_{p_2}, \dots, F'_{p_k}))) \end{aligned} \tag{11}$$

3.2.3. Multi-scale fusion (MSF) bridge

The success of U-Net in medical image field is attributed to the effective fusion of coarse-grained and fine-grained features between encoder and decoder, which is crucial for restoring details in dense segmentation. To compensate for the detail loss from the concise polar token generation process, this study applies a multi-scale fusion bridge (Fig. 7) to generate comprehensive fused features for the subsequent encoder and decoder skipping connection. The concise polar features from DCCA blocks in the encoding phase $\widetilde{F_p}^i (i = 1, 2, \dots, L)$ are flattened, concatenated and fed into a multi-head self-attention (MHSA) module. Afterward, the fused features are split and reshaped to the original shape.

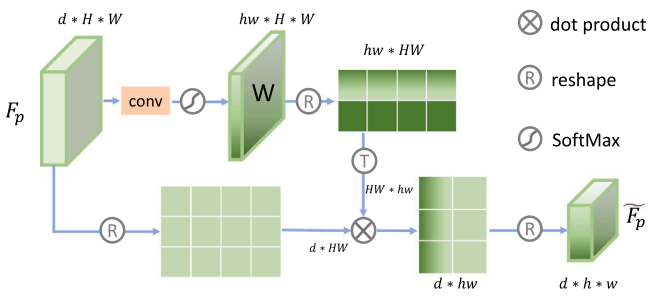

Fig. 5. Schematic illustration of self-attention (SA) module applied on polar features to generate concise tokens.

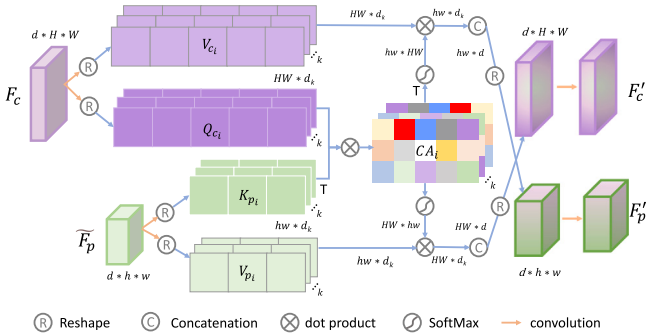

Fig. 6. Schematic diagram of the proposed multi-head cross-attention (MHCA) module.

3.2.4. Explicit position encoding

To facilitate position encoding for Transformer, we generate position maps for the purpose of explicit position encoding. In the Cartesian coordinate, the map is generated by normalizing the Euclidean distance of each pixel to the image center within the range of 0–1. Subsequently, the Cartesian map is transformed into log-polar or linear-polar coordinate following the same formula as the input image transformation. The dual-coordinate images are concatenated with their corresponding maps in the channel axis to form the input. The visualization of the position maps is shown in supplementary Fig. A1.

4. Experiments

4.1. Coronary OCT data set

In the model development phase, all OCT image pullbacks at the CardHemo core lab (Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China) are screened. A total of 395 OCT pullbacks from 339 patients with thrombus are included, resulting in 5649 OCT cross-sections. The data is split into training and validating sets with a ratio of 4:1 at the patient level for learning parameters and tuning hyperparameters, respectively. After that, the final evaluation of models' performance is conducted on an independent data set from Oxford Acute Myocardial Infarction (Ox-AMI) study (REC number 10/H0408/24, Oxford Heart Center, UK), which consists of 548 cross-sections from 79 OCT pullbacks of 52 patients. All evaluation metrics

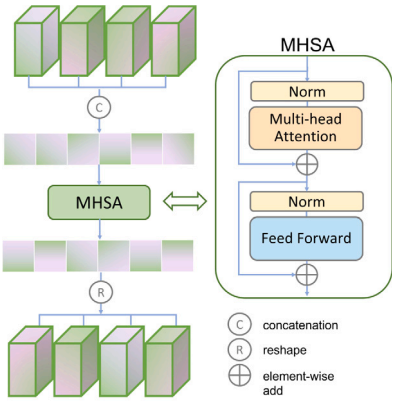

Fig. 7. Schematic diagram of the multi-scale fusion bridge implemented by multi-head self-attention mechanism.

reported in this section are based on this external testing data, which is not used during the model development phase. All OCT images are acquired in the clinical routine using OCT systems of C7-XR™ or ILUMIEN OPTIS™ FD-OCT system with Dragonfly™ catheters (Abbott, USA). Thrombus clot, guide wire and flow area are manually labeled by two experienced OCT analysts (MC, SB) to generate ground truths. All patients provided informed consent for enrollment in the institutional database for potential future investigations.

4.2. Evaluation metrics

Evaluation metrics of Dice Similarity Coefficient (DSC) and Hausdorff Distance 95% percentile (HD95) are used to report and compare performance across different models.

$$DSC = \frac{2 \times TP}{2 \times TP + FN + FP}$$
$$HD = max\{\max_{s_P \in S(P)} d(s_P, S(M)), \max_{s_M \in S(M)} d(s_M, S(P))\} \quad (12)$$

where TP, FN and FP are true positive, false negative and false positive, respectively. $d(s_P, S(M))$ is the shortest Euclidean distance from the prediction pixel s_P to the mask boundary set $S(M)$. Similarly, $d(s_M, S(P))$ is the shortest Euclidean distance from mask pixel s_M to the prediction boundary set $S(P)$.

4.3. Statistical analysis

Normal distribution of continuous variables is tested by Shapiro–Wilk test. Continuous variables with normal distribution are presented using mean value, otherwise the median value is reported. Comparison of continuous variables is performed with t-test or nonparametric Mann–Whitney's test, depending on the distribution of data. Statistical significance is set at the 0.05 level. All statistical analyses are performed with SPSS 25 (SPSS Inc., Chicago, IL).

4.4. Implementation details

All models are implemented on NVIDIA A100 GPU using Pytorch platform (Paszke et al., 2019). Models are initialized randomly and trained from scratch, apart from Swin-Unet. SGD is used as the optimizer with momentum 0.9 and weight decay of 5e−4. Warm-Up learning rate strategy is employed with an initial learning rate of 5e−3. Maximum training epochs are set to 600 with a batch size of 32. The best model is selected based on the validating data set and then is compared using the independent external testing data set.

4.4.1. Loss function

The designed model conducts a multi-class segmentation task, where target objects of thrombus clot and guide wire occupy a small portion of the whole image, leading to a class imbalance challenge. Following our previous work (Chu et al., 2021), focal Tversky loss Eq. (13), which is considered a generalization of Dice loss, is adopted to tune trade-offs between false negatives and false positives via hyperparameters of $\alpha = 0.7$ and $\beta = 0.3$. Hyper-parameter γ allows the model to focus on poorly performed categories and is set to 0.75, as previously suggested (Abraham and Khan, 2019). Moreover, to stabilize the training process, pixel-wise cross-entropy loss Eq. (14) is added to the total loss Eq. (15) with a magnitude balance $\lambda = 2$.

$$Loss_{FT} = \sum_{c=1}^{C} (1 - \frac{TP_c}{TP_c + \alpha \cdot FN_c + \beta \cdot FP_c})^{\gamma} \quad (13)$$

$$Loss_{CE} = -\frac{1}{N} \sum_{c=1}^{C} \sum_{n=1}^{N} g_{nc} log p_{nc} \quad (14)$$

$$Loss = Loss_{FT} + \lambda Loss_{CE} \quad (15)$$

C represents the total classes of target objects and N represents total pixel numbers of the input image; TP_c, FN_c and FP_c are true positives, false negatives and false positives for object c; p_{nc} and g_{nc} are the prediction and ground truth values for pixel n of object c.

4.4.2. Data augmentation

On-the-fly data augmentation is conducted for all models to increase data diversity. For both coordinates, random contrast, brightness, gamma transformation, vertical and horizontal flipping, and grid distortion are applied. For Cartesian images, shifting, scaling and rotation are conducted to add geometric diversity, whilst the abovementioned augmentation is replaced by a pixel dropout for polar images.

5. Experimental results

5.1. Segmentation performance compared with other methods

We design a series of comparative studies to validate the superiority of our method over other methods, including both CNN-based and Transformer-based models. Table 1 provides quantitative comparison results of these two groups. Overall, the proposed model exhibits distinguished performance compared to all other methods, achieving the highest DSC and lowest HD95 values across all objects. The Mann–Whitney's test confirms the statistical significance of this superiority, with a p-value < 0.01 in most cases. When comparing the two method groups, CNN-based methods generally outperform Transformer-based methods. Figs. 8 and 9 visualize the comparison between the proposed model and CNN-based and Transformer-based models, respectively.

Regarding the main target of thrombus segmentation, U-Net and its variants significantly outperform DeepLabv3+ in the CNN-based group, underscoring the importance of sophisticated decoding paths for achieving precise segmentation in medical images. Although U-Net and its variants achieve comparable DSC, the HD95 values are better for UNet++ and Attention U-Net, indicating more accurate segmentation on the boundaries. In the Transformer-based group, Swin-Unet is an Unet-like pure Transformer model while UNETR and TransUNet are hybrid networks combining CNN and Transformer. Swin-Unet is initialized with pre-trained weights from ImageNet, as provided in the original publication due to poor performance observed when training from scratch. Among these models, UNETR achieves the best thrombus segmentation performance. TransUNet has a slightly higher DSC yet larger HD95 than Swin-Unet, indicating no absolute superiority of this hybrid design over the pure Transformer design. This is likely because the attention layers in TransUNet are only added at the highest level of semantic features, limiting comprehensive feature extraction at the shallow levels which contain rich details. As observed

Table 1
Comparison of DSC and HD95 for the proposed model with other developed methods on external testing data.

		Average		Flow area		Thrombus clot		Guide wire	
		DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
CNN-based	U-Net	0.738**	0.305**	0.950*	0.117**	0.656**	0.510**	0.662**	0.064**
	UNet++	0.743**	0.288**	0.948**	0.117**	0.653**	0.499**	0.707**	0.055**
	Attention U-Net	0.732**	0.302**	0.949*	0.117**	0.655**	0.487**	0.667**	0.056**
	DeepLabv3+	0.642**	0.513**	0.938**	0.156**	0.584**	0.591**	0.533**	0.361**
Transformer-based	UNETR	0.578**	0.557**	0.930**	0.175**	0.404**	0.751**	0.521**	0.309**
	TransUNet	0.574**	0.614**	0.922**	0.196**	0.375**	1.062**	0.496**	0.121**
	Swin-Unet[a]	0.472**	0.737**	0.909**	0.274**	0.361**	0.864**	0.170**	0.910**
Proposed		**0.784**	**0.222**	**0.951**	**0.087**	**0.706**	**0.399**	**0.732**	**0.050**

DSC: Dice Similarity Coefficient; HD95: 95% Hausdorff Distance (mm); ↓: Lower is better; ↑: Higher is better. Significant difference with the proposed model is indicated by * and **, when p-value < 0.05 and < 0.01, respectively.

[a] Indicates the model is initialized with pre-trained weights on ImageNet.

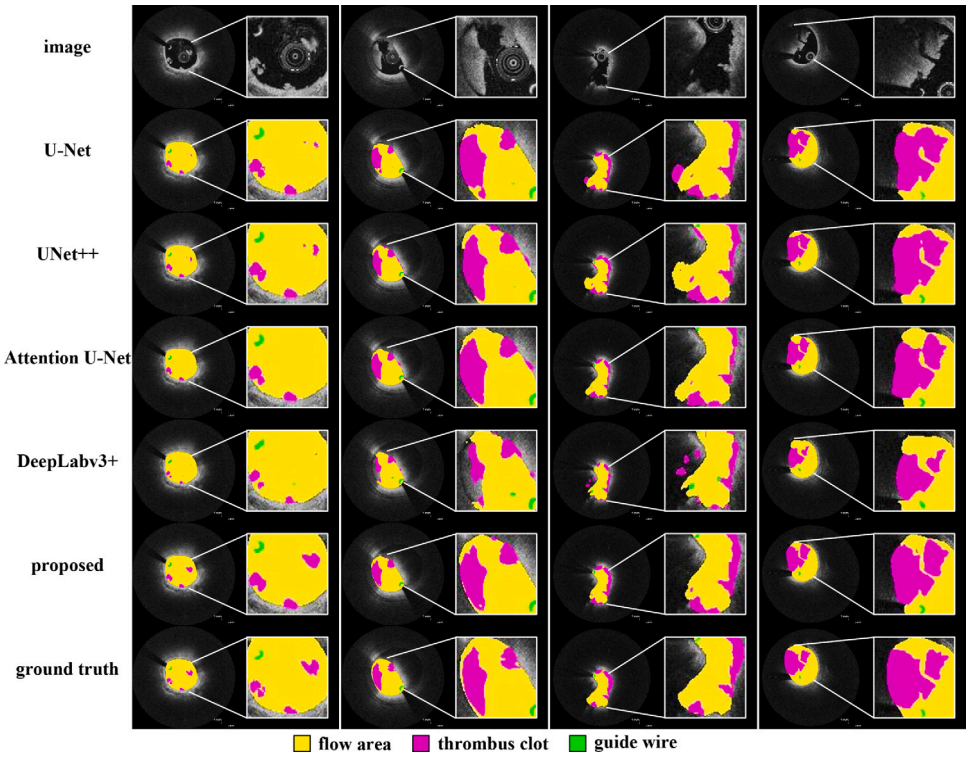

Fig. 8. Visualization comparisons between CNN-based models and the proposed model. Rows from top to bottom stand for input images, predictions by different methods and the ground truths.

in Fig. 9, the segmentation results by TransUNet are over-smoothed. This highlights the importance of comprehensive combination of the attentional and convolutional layers across multiple scales for hybrid networks. By leveraging dual-coordinate image features via the cross-attention mechanism hierarchically, the proposed model outperforms other Transformer-based models on thrombus segmentation by a large margin.

5.2. Robustness to geometric transformation

To assess the geometric robustness attributed to the additional polar images, four sets of random data augmentation (scaling, rotate, shift, and their combination) are applied to the original external testing data, generating four different augmented testing sets. Geometric transformation is first applied to the Cartesian image, which is subsequently projected into the corresponding Polar image.

UNet++ and UNETR, chosen as the best representatives of CNN-based and Transformed-based models, are retested and compared with the proposed model on the augmented data. The changes in DSC value across different augmentations are depicted in a heatmap (Fig. 10). Integration of polar images through the proposed DCCA block leads to more consistent predictions, resulting in smaller DSC changes in comparison to UNet++ and UNETR. Fig. 11 gives a qualitative visualization of models' response to different augmentations.

5.3. Date efficiency

The need of large-scale training data for Transformer models poses a challenge for their application in medical imaging. Data efficiency, which refers to a model's ability to achieve high performance with relatively small amounts of training data, is critical in this context. We investigate the data efficiency of our proposed model by evaluating its performance under low-data regimes (10%, 25%, and 50% of the total training data). Fig. 12 shows the performance comparison of our model with other Transformer models at different percentages of training data. As the training data decreases, our model achieves consistently higher DSC and lower HD95 values compared to other methods. Notably, the proposed model trained with 10% of total training data shows

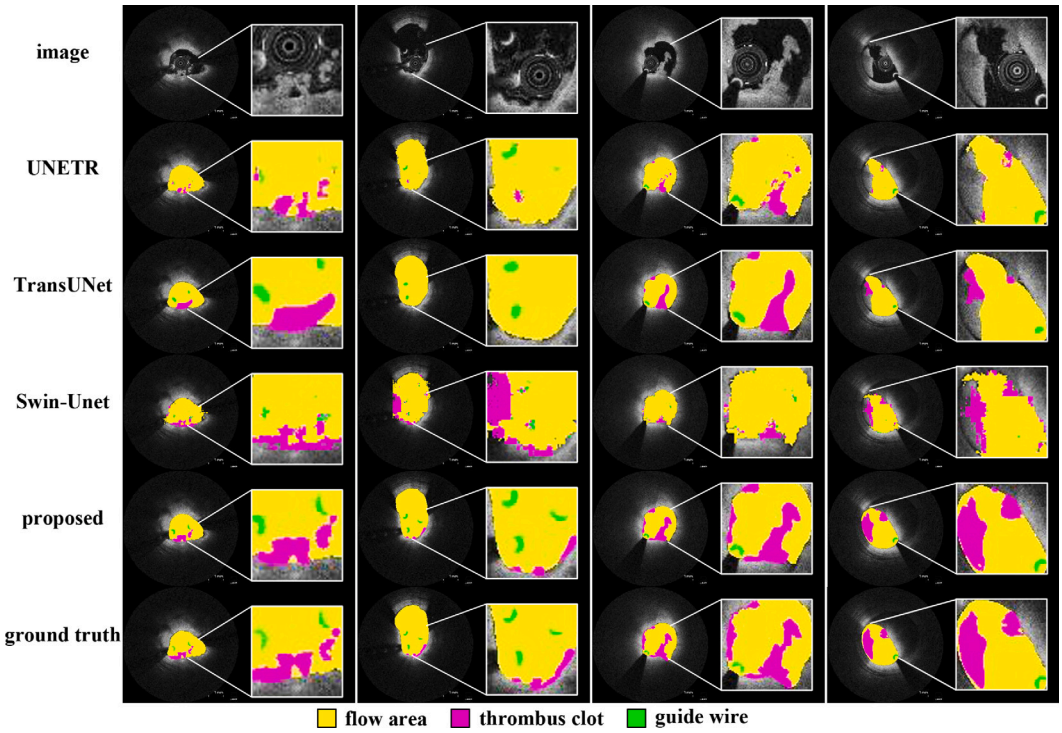

Fig. 9. Visualization comparisons between Transformer-based models and the proposed model. Rows from top to bottom stand for input images, predictions by different methods and ground truths. Notably, the cases shown are independent from Fig. 8.

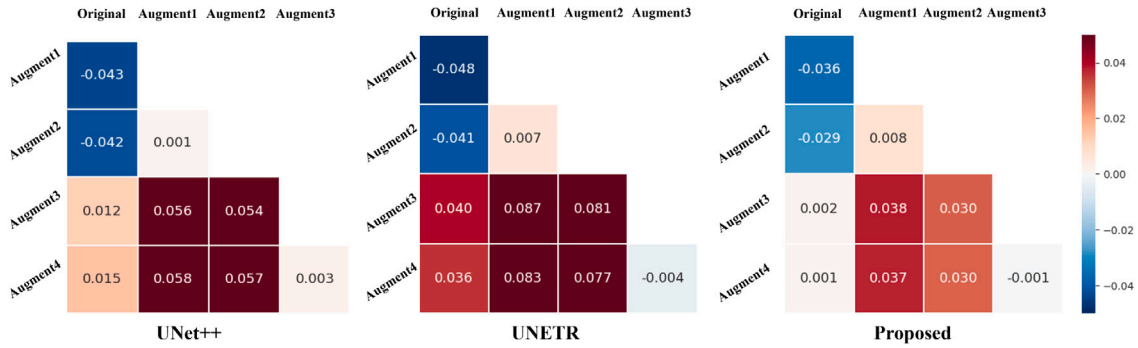

Fig. 10. Heatmap of DSC changes for UNet++, UNETR, and the proposed model among four augmented testing data sets. Augment1: combination of shift, scaling and rotate; Augment2: scaling; Augment3: rotate; Augment4: shift.

comparable performance to UNETR or TransUNet using full data, which proves the superior data efficiency of our method.

5.4. Ablation study

We first compare the performance of log-polar and linear-polar OCT transformation. As shown in Table 2, dual-coordinate input with linear-polar image shows a better performance than log-polar image in terms of average DSC and HD95. For specific objects, log-polar outperforms linear-polar in segmenting thrombus (DSC: 0.718 vs. 0.706), potentially because the log-polar transformation encodes both rotation and scaling into shift transformation, whereas linear-polar transformation only encodes rotation. Nevertheless, considering the overall performance and the fact that the raw data of OCT acquired by the transducer is transformed to Cartesian image in a linear mapping to reserve pixel spacing, we choose linear-polar image as the proposed method.

We then perform ablation studies to verify the contribution of each essential component by conducting the following experimental settings: (1) The proposed model structure utilizing a single Cartesian image as input without cross-attention, is referred to as Ablation1; (2) Ablation2 uses polar image instead of Cartesian image as input under the same architecture; Then the proposed DCCA block is added to the model with dual-coordinate images, with either (3)Multi-scale fusion (MSF) bridge, referred as Ablation3, or (4)Explicit position encoding (EPE), referred as Ablation4; Lastly, both components of MSF and EPE are added to the model (Proposed). All models are tuned to be in the same magnitude order of parameters and FLOPs via adjusting channels of convolutional kernel. The motivation is to make sure performance differences are attributed to the designed component rather than the amount of model parameters.

As shown in Table 3, models with one view input of Cartesian or polar image achieve comparable performance on average DSC (0.741 vs. 0.748). However, the model with polar input has better performance than with Cartesian input on guide wire (DSC 0.683 vs. 0.732) at the cost of a performance drop in thrombus (DSC: 0.663 vs. 0.631), which is in line with advantageous features present in each coordinate. The performance drop on thrombus can be compensated by adding DCCA blocks. Notably, both MSF and EPE bring benefits to the model

Table 2
Performance comparisons of log-polar and linear-polar OCT transformation.

View	Average		Flow area		Thrombus clot		Guide wire	
	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
Cartesian+Log-polar	0.774	0.259	**0.952**	**0.087**	**0.718**	0.445	0.724**	**0.042**
Cartesian+Linear-polar	**0.784**	**0.222**	0.951	**0.087**	0.706	**0.399**	**0.732**	0.050

Significant difference between the two models is indicated by ** when p-value < 0.01. Other abbreviations are the same as Table 1.

Table 3
Performance comparisons of ablation studies.

Num	Design				Scale		Average		Flow area		Thrombus clot		Guide wire	
	View	DCCA	MSF	EPE	FLOPs(G)	Params(M)	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
1	C				41.62	39.21	0.741	0.274	0.948	0.111	0.663	0.469	0.683	0.055
2	P				41.62	39.21	0.748	0.593	**0.952**	0.196	0.631	1.151	**0.732**	0.087
3	C+P	√	√		38.85	40.23	0.771	0.249	0.949	0.111	0.693	0.446	0.723	**0.049**
4	C+P	√		√	38.83	36.16	0.766	0.265	0.949	0.111	0.681	0.472	0.726	0.052
Proposed	C+P	√	√	√	38.89	40.23	**0.784**	**0.222**	0.951	**0.087**	**0.706**	**0.399**	**0.732**	0.050

C: Cartesian image; P: linear-polar image; DCCA: dual-coordinate cross-attention; MSF: multi-scale fusion; EPE: explicit position encoding. Other abbreviations are the same as Table 1.

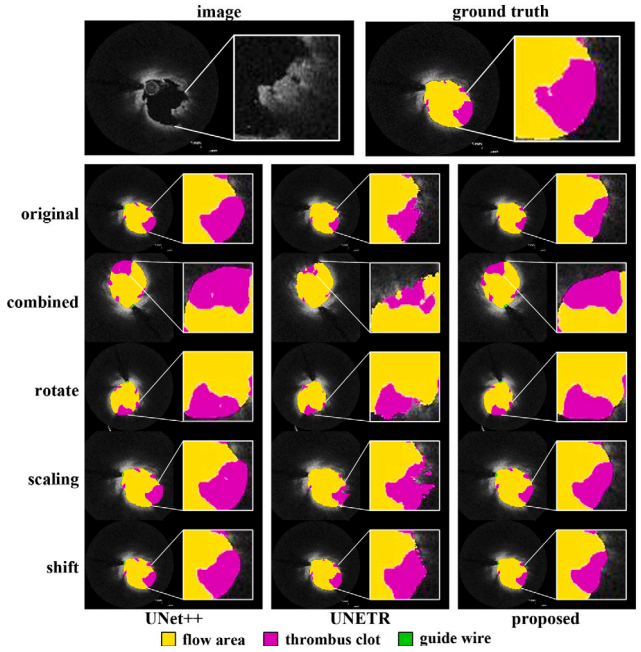

Fig. 11. Visualization of models' predictions on four augmented testing data. The first row shows the image and the ground truth. The following rows stand for models' predictions on the original image and four augmented images. The columns from left to right are predictions by UNet++, UNETR and the proposed model.

performance, with the average DSC increasing from 0.741 to 0.771 and 0.766, respectively.

Fig. 13 visualizes the deepest features of thrombus through gradient-weighted class activation mapping for models with either sole Cartesian (Fig. 13A) or polar image (Fig. 13B), in comparison with dual-coordinate images (Fig. 13C–D). By leveraging advantageous features from dual-coordinate images, relevant features are enhanced by each other and contribute more significantly to the final prediction. Meanwhile, distraction from surrounding irrelevant noises is suppressed, resulting in a sharper boundary.

We also demonstrate the effectiveness of the combined focal Tversky loss and CE loss by comparing it with other loss functions: (1) a combined Dice and CE loss, which is commonly used in segmentation tasks; (2) a combined focal Dice and CE loss by setting the false positives and false negatives hyper-parameters in focal Tversky loss to $\alpha = 0.5$ and $\beta = 0.5$. As shown in Table 4, a significant DSC drop is observed when the focal Tversky loss is replaced by Dice loss or focal Dice loss, particularly for the thrombus, indicating the effectiveness of the combined loss used in the study.

5.5. Transferability study

The proposed dual-coordinate cross-attention design can easily cooperate with other developed models. We replace the original self-attention block in UNETR and TransUNet with the proposed multi-head DCCA block. The structure modifications of these mature models are straightforward, following the principle of their initial designs. Schematic illustrations of the modification to the abovementioned models are presented in Supplementary Fig. A2–A3. The diagram style is consistent with the original study to facilitate comparison by readers.

As shown in Table 5, after replacing self-attention within the single image with cross-attention across dual-coordinate images, the segmentation performance has significantly improved, especially for the challenging object of thrombus. Fig. 14 provides visualizations of the comparison. The extremely deviated shape of thrombus predicted by UNETR and TransUNet is alleviated after adding DCCA. Moreover, the mosaic effect observed in UNETR has also been improved, leading to smoother boundaries.

Fig. 15 visualizes the deepest features of Transformer models w/o DCCA block for segmenting thrombus. The original model tends to be disturbed by noises and overlook the target. Dual-coordinate cross-attention helps the model focus on the right contexts with higher gradients, leading to better performances.

6. Discussion and conclusions

To the best of our knowledge, this is the first study investigating deep learning method to automatically segment thrombus on coronary OCT for patients with acute coronary syndromes. The challenge lies in the complex imaging features of OCT where both global context and local details are crucial for precise segmentation. Additionally, the limited amount of clinical OCT data necessitates the use of data-efficient DL networks. To address these challenges, we propose a novel dual-coordinate cross-attention transformer (DCCAT) network that enhances relevant features from both Cartesian and Polar coordinates through long-range dependency modeling. The superiority, robustness and data efficiency of DCCAT are evaluated using independent external testing data. The proposed dual-coordinate cross-attention design can be flexibly plugged into other developed Transformer models to boost performance.

The novelty of the current model lies in several aspects. Firstly, the inputs of the model are the same OCT image from two coordinates. Polar image is more advantageous for guide wire segmentation because

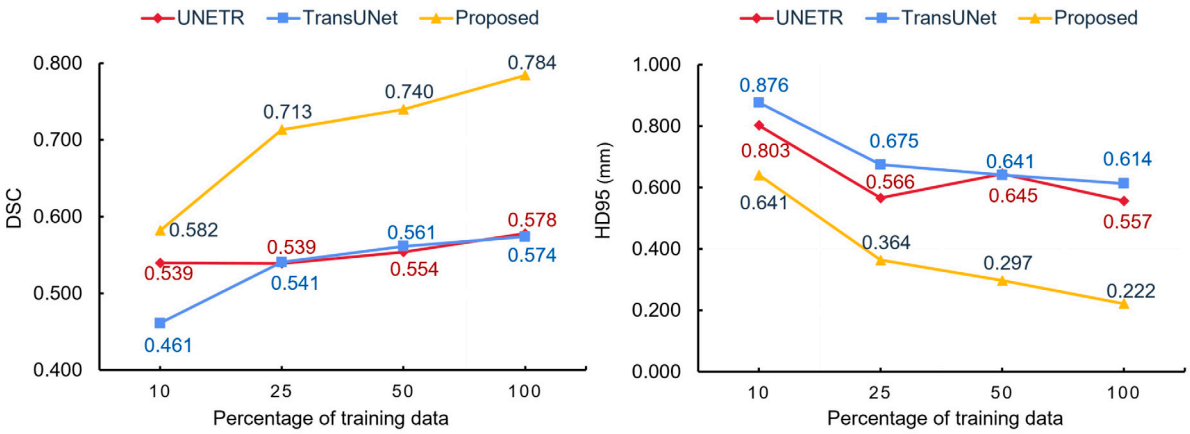

Fig. 12. DSC and HD95 comparison on the external testing data at different percentages of training samples.

Table 4
Performance comparisons of loss function.

Loss	Average		Flow area		Thrombus clot		Guide wire	
	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
Dice+CE	0.767**	0.291**	0.950*	0.111**	0.672**	0.494**	0.732	**0.039***
Focal Dice+CE	0.764**	0.249**	0.950	0.096**	0.677**	0.467**	0.720**	**0.039**
Proposed	**0.784**	**0.222**	**0.951**	**0.087**	**0.706**	**0.399**	**0.732**	0.050

Significant difference with the proposed model is indicated by * and **, when p-value < 0.05 and < 0.01, respectively. Other abbreviations are the same as Table 1.

Table 5
Effectiveness of the DCCA block embedded in other Transformer models.

	Average		Flow area		Thrombus clot		Guide wire	
	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓	DSC↑	HD95↓
UNETR	0.578	0.557	0.930	0.175	0.404	0.751	0.521	0.309
UNETR+DCCA	**0.623****	**0.487****	**0.937****	**0.141****	**0.483****	**0.718**	**0.550****	**0.166****
TransUNet	0.574	0.614	0.922	0.196	0.375	1.062	0.496	0.121
TransUNet+DCCA	**0.592****	**0.583****	0.922	0.196**	**0.412****	**1.017****	**0.538****	**0.115****

Significant difference with a p-value < 0.05 and < 0.01 is indicated by * and **, respectively. Other abbreviations are the same as Table 1.

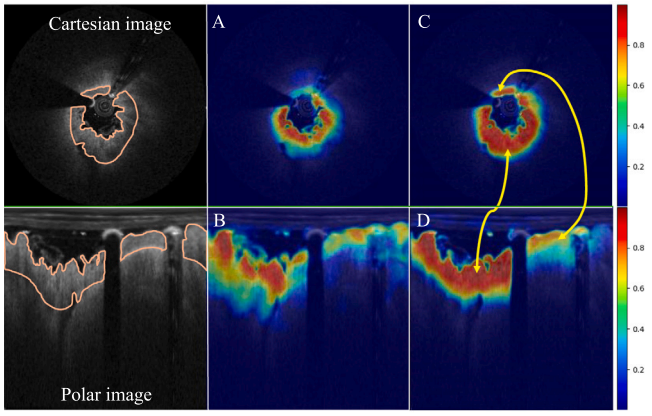

Fig. 13. Visualization of the deepest feature maps of thrombus for models with (A) single Cartesian image input, (B) single polar image input, and (C–D) dual-coordinate images. The orange contours superimposed on the left column denote thrombus clots. Yellow arrows on the right column indicate the association between regions across dual-coordinate images.

of the distinctive signal attenuation behind it. On the other hand, the elliptical shape of the coronary lumen in Cartesian OCT makes identifying thrombus relatively easier. Despite the discriminative features in each coordinate, thrombus and guide wire share similarities in texture, shape, and brightness in 2D dimensional space, with thrombus clots appearing lumpy and guide wires crescent shaped. There may be other relevance in higher spatial dimensions extracted by the network that is beyond human vision. We exploit the complementarity and similarity of the two coordinate OCT images through multi-view image DL. By leveraging features from dual-coordinate images, relevant features are enhanced by each other and contribute more significantly to the final prediction. Our experiments demonstrate that dual-coordinate images lead to enhanced semantic understanding and boosted performance, especially in segmenting thrombus and guide wire.

Secondly, the additional polar image improves the model's robustness to geometric transformations by leveraging the shift equivalence property of CNN. The combination of polar transformation and CNN functions as a warped convolution, extending the inherent shift equivalence property of convolutional operations to other geometric transformations. This approach addresses the lack of inductive bias in Transformers, improving data efficiency and geometric robustness. As demonstrated in our experiments, the proposed model trained with 10% data achieves comparable performance to UNETR or TransUNet using full data. The performances remain more consistent across various geometric augmentations.

Lastly, feature fusion from the two views is implemented by cross-attention mechanism in a hierarchical way. Efficient feature fusion plays a key role in multi-view image DL. Previous studies usually fuse features only at the highest-level, which constrains feature communications at lower layers. Differently, we stack the proposed dual-coordinate cross-attention block amid convolutional layers at multiple

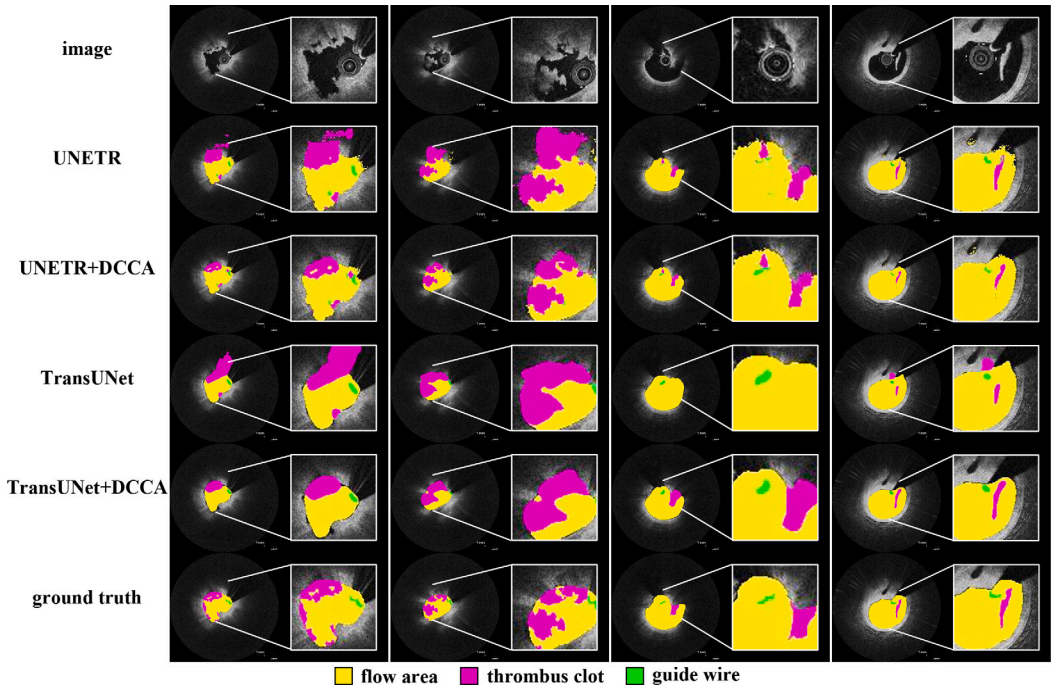

Fig. 14. Visualization comparison of Transformer models w/o DCCA block. The rows from top to end stand for images, models' predictions, and the corresponding ground truths.

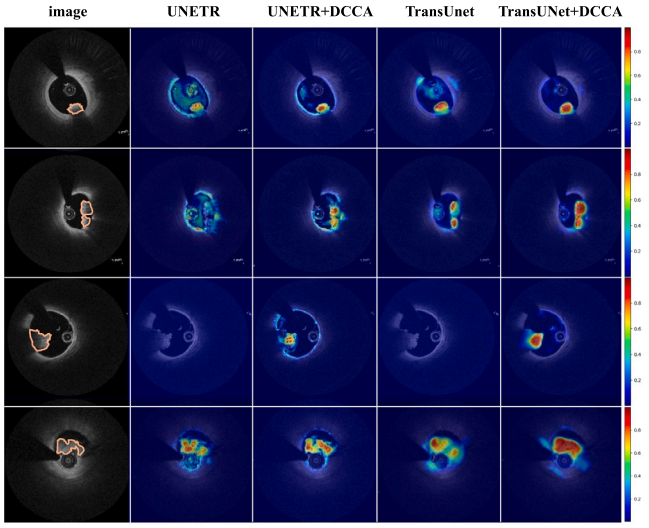

Fig. 15. Visualization of the deepest feature maps of thrombus clots for Transformer models w/o DCCA. Columns from left to right stand for input images, UNETR w/o DCCA and TranUNet w/o DCCA. The orange contours superimposed on the left column denote thrombus clots.

levels through the model, enabling comprehensive feature enhancement and complementarity at various scales. To alleviate the computational complexity introduced by the attention layers, only one pair of Query and Key embeddings are used to mine long-range dependencies. Meanwhile, concise token maps are adopted to ensure a linear computational complexity.

In recent years, DL has facilitated diagnosis and prognosis in cardiovascular imaging, especially for patients with chronic coronary syndromes (Chu et al., 2023). Nevertheless, the potential of DL in emergent scenarios is worthy of further investigation to bring clinical benefits, considering its fast analysis speed and the necessity of rapid decision-making. It is important to note that the analysis speed for the proposed model is 2–3 s per pullback during the inference period on NVIDIA RTX A4000, whereas manual annotation by analysts can take several hours. The model can be used for real-time analysis in the cathlab and exhibits promising clinical applications. For ACS patients with large thrombus burden, aspiration thrombectomy before stenting can reduce microvascular injury caused by thromboembolism in theory. However, large randomized controlled trials have shown controversial results (Fröbert et al., 2013; Jolly et al., 2015). A potential reason is the lack of effectiveness evaluation on the aspiration thrombectomy before proceeding to stenting. The proposed model allows for automatic quantification of thrombus, which provides a tool to evaluate the efficacy of thrombectomy and identify subgroups of patients who would benefit most from aspiration thrombectomy. Moreover, anti-thrombotic therapy without stenting in selected ACS patients caused by plaque erosion has been proven sufficient to restore coronary artery patency. In comparison to plaque rupture, plaque erosion has an intact fibrous cap, larger lumen and smaller thrombus burden. The model could aid in distinguishing different ACS mechanisms, and therefore help clinical strategy management. Additionally, the amount of residual thrombus has been shown to correlate with clinical outcomes (Zhou et al., 2021). Future work will explore the assessment of residual thrombus after stenting, which may be useful for a patient-level risk-stratification. The model considers all thrombi as one category without distinguishing between white or red phenotypes due to the common mixed existence and the lack of standards for differentiation by visual assessment. We anticipate that with more advanced imaging techniques or computer-aided quantification (Kaivosoja et al., 2018; De Maria et al., 2017), distinguishing between white and red thrombus will become more accurate.

In conclusion, this study presents a novel deep learning model for automatic thrombus segmentation on coronary OCT for the first time. Its data efficiency, robust performance, and real-time analysis speed make it a promising tool for clinical use.

CRediT authorship contribution statement

Miao Chu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Giovanni Luigi De Maria:** Writing – review

& editing, Supervision, Resources, Project administration, Data curation, Conceptualization. **Ruobing Dai:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis. **Stefano Benenati:** Writing – review & editing, Visualization, Formal analysis, Data curation. **Wei Yu:** Writing – review & editing, Software, Data curation. **Jiaxin Zhong:** Methodology, Data curation. **Rafail Kotronias:** Resources, Data curation. **Jason Walsh:** Writing – review & editing, Visualization. **Stefano Andreaggi:** Writing – review & editing, Validation. **Vittorio Zuccarelli:** Validation, Writing – review & editing. **Jason Chai:** Writing – review & editing, Data curation. **Oxford Acute Myocardial Infarction (OxAMI) Study investigators:** Data curation. **Keith Channon:** Writing – review & editing, Supervision, Resources. **Adrian Banning:** Writing – review & editing, Supervision, Resources, Conceptualization. **Shengxian Tu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

S Tu reported research grants and consultancy from Pulse Medical. A Banning reports institutional research grant from Boston Scientific. G De Maria reports research grant from Miracor, Medtronic, Terumo, Abbott, Philips and consultant fee from Miracor. R Kotronias, K Channon, A Banning, and G De Maria acknowledge support/funding from the Oxford NIHR Biomedical Research center. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose

Data availability

The data that has been used is confidential.

Acknowledgments

We acknowledge Prof. Wei Yang (Southern Medical University) for his insightful advice on the manuscript of this paper.

Fundings

This study is supported by the National Natural Science Foundation of China (82327808 to ST and 82302285 to MC), the China Scholarship Council, China and Medical-engineering Research Project, China (YG2023QNA03) to MC.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103265.

References

Abraham, N., Khan, N.M., 2019. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 683–687.

Ahn, S.S., Ta, K., Thorn, S.L., Onofrey, J.A., Melvinsdottir, I.H., Lee, S., Langdon, J., Sinusas, A.J., Duncan, J.S., 2023. Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3D echocardiography. Med. Image Anal. 84, 102711.

Aleong, G., Vaqueriza, D., Del Valle, R., Garcia, H., Hernandez, R., Alfonso, F., Jimenez-Quevedo, P., Bañuelos, C., Macaya, C., Escaned, J., 2009. Dual quantitative coronary angiography: a novel approach to quantify intracoronary thrombotic burden. Eurointervention 4 (4), 475–480.

Athanasiou, L.S., Bourantas, C.V., Rigas, G., Sakellarios, A.I., Exarchos, T.P., Siogkas, P.K., Ricciardi, A., Naka, K.K., Papafaklis, M.I., Michalis, L.K., et al., 2014. Methodology for fully automated segmentation and plaque characterization in intracoronary optical coherence tomography images. J. Biomed. Opt. 19 (2), 026009–026009.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2023. Advances in medical image analysis with vision transformers: a comprehensive review. Med. Image Anal. 103000.

Chen, C.-F.R., Fan, Q., Panda, R., 2021a. Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366.

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

Chen, J., Luo, Z., Zhang, Z., Huang, F., Ye, Z., Takiguchi, T., Hancock, E.R., 2018a. Polar transformation on image features for orientation-invariant representations. IEEE Trans. Multimed. 21 (2), 300–313.

Chu, M., Jia, H., Gutierrez-Chico, J.L., Maehara, A., Ali, Z.A., Zeng, X., He, L., Zhao, C., Matsumura, M., Wu, P., et al., 2021. Artificial intelligence and optical coherence tomography for the automatic characterisation of human atherosclerotic plaques. EuroIntervention 17 (1), 41–50.

Chu, M., Wu, P., Li, G., Yang, W., Gutiérrez-Chico, J.L., Tu, S., 2023. Advances in diagnosis, therapy, and prognosis of coronary artery disease powered by deep learning algorithms. JACC: Asia 3 (1), 1–14.

Cohen, T., Welling, M., 2016. Group equivariant convolutional networks. In: International Conference on Machine Learning. PMLR, pp. 2990–2999.

Crea, F., Libby, P., 2017. Acute coronary syndromes: the way forward from mechanisms to precision treatment. Circulation 136 (12), 1155–1166.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

De Maria, G.L., Patel, N., Wolfrum, M., Fahrni, G., Kassimis, G., Porto, I., Dawkins, S., Choudhury, R.P., Forfar, J.C., Prendergast, B.D., et al., 2017. The influence of coronary plaque morphology assessed by optical coherence tomography on final microvascular function after stenting in patients with ST-elevation myocardial infarction. Coron. Artery Dis. 28 (3), 198–208.

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 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y., 2018. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 264–272.

Fröbert, O., Lagerqvist, B., Olivecrona, G.K., Omerovic, E., Gudnason, T., Maeng, M., Aasa, M., Angerås, O., Calais, F., Danielewicz, M., et al., 2013. Thrombus aspiration during ST-segment elevation myocardial infarction. New Engl. J. Med. 369 (17), 1587–1597.

Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N., 2022. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131.

Gessert, N., Heyder, M., Latus, S., Lutz, M., Schlaefer, A., 2018a. Plaque classification in coronary arteries from ivoct images using convolutional neural networks and transfer learning. arXiv preprint arXiv:1804.03904.

Gessert, N., Lutz, M., Heyder, M., Latus, S., Leistner, D.M., Abdelwahed, Y.S., Schlaefer, A., 2018b. Automatic plaque detection in IVOCT pullbacks using convolutional neural networks. IEEE Trans. Med. Imaging 38 (2), 426–434.

Giacoppo, D., Laudani, C., Occhipinti, G., Spagnolo, M., Greco, A., Rochira, C., Agnello, F., Landolina, D., Mauro, M.S., Finocchiaro, S., et al., 2024. Coronary angiography, intravascular ultrasound, and optical coherence tomography for guiding of percutaneous coronary intervention: a systematic review and network meta-analysis. Circulation 149 (14), 1065–1086.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

Henriques, J.F., Vedaldi, A., 2017. Warped convolutions: Efficient invariance to spatial transformations. In: International Conference on Machine Learning. PMLR, pp. 1461–1469.

Holm, N.R., Andreasen, L.N., Neghabat, O., Laanmets, P., Kumsars, I., Bennett, J., Olsen, N.T., Odenstedt, J., Hoffmann, P., Dens, J., et al., 2023. OCT or angiography guidance for PCI in complex bifurcation lesions. New Engl. J. Med. 389 (16), 1477–1487.

Huang, Y., He, C., Wang, J., Miao, Y., Zhu, T., Zhou, P., Li, Z., 2018. Intravascular optical coherence tomography image segmentation based on support vector machine algorithm. MCB Mol. Cell. Biomech. 15 (2), 117–125.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

Jia, H., Abtahian, F., Aguirre, A.D., Lee, S., Chia, S., Lowe, H., Kato, K., Yonetsu, T., Vergallo, R., Hu, S., et al., 2013. In vivo diagnosis of plaque erosion and calcified nodule in patients with acute coronary syndrome by intravascular optical coherence tomography. J. Am. Coll. Cardiol. 62 (19), 1748–1758.

Jia, H., Dai, J., Hou, J., Xing, L., Ma, L., Liu, H., Xu, M., Yao, Y., Hu, S., Yamamoto, E., et al., 2017. Effective anti-thrombotic therapy without stenting: intravascular optical coherence tomography-based management in plaque erosion (the EROSION study). Eur. Heart J. 38 (11), 792–800.

Jolly, S.S., Cairns, J.A., Yusuf, S., Meeks, B., Pogue, J., Rokoss, M.J., Kedev, S., Thabane, L., Stankovic, G., Moreno, R., et al., 2015. Randomized trial of primary PCI with or without routine manual thrombectomy. New Engl. J. Med. 372 (15), 1389–1398.

Kaivosoja, T.P., Liu, S., Dijkstra, J., Huhtala, H., Sheth, T., Kajander, O.A., 2018. Comparison of visual assessment and computer image analysis of intracoronary thrombus type by optical coherence tomography. PLoS One 13 (12), e0209110.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N., 2020. Big transfer (bit): General visual representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, pp. 491–507.

Lei, B., Huang, S., Li, H., Li, R., Bian, C., Chou, Y.-H., Qin, J., Zhou, P., Gong, X., Cheng, J.-Z., 2020. Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Med. Image Anal. 64, 101753.

Li, C., Jia, H., Tian, J., He, C., Lu, F., Li, K., Gong, Y., Hu, S., Yu, B., Wang, Z., 2021. Comprehensive assessment of coronary calcification in intravascular oct using a spatial-temporal encoder-decoder network. IEEE Trans. Med. Imaging 41 (4), 857–868.

Liu, D., Gao, Y., Zhangli, Q., Han, L., He, X., Xia, Z., Wen, S., Chang, Q., Yan, Z., Zhou, M., et al., 2022. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 485–495.

Lu, X., Wang, W., Shen, J., Crandall, D., Luo, J., 2020. Zero-shot video object segmentation with co-attention siamese networks. IEEE Trans. Pattern Anal. Mach. Intell. 44 (4), 2228–2242.

Park, S., Araki, M., Nakajima, A., Lee, H., Fuster, V., Ye, J.C., Jang, I.-K., 2022. Enhanced diagnosis of plaque erosion by deep learning in patients with acute coronary syndromes. Cardiovasc. Interv. 15 (20), 2020–2031.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Porto, I., Mattesini, A., Valente, S., Prati, F., Crea, F., Bolognese, L., 2015. Optical coherence tomography assessment and quantification of intracoronary thrombus: Status and perspectives. Cardiovasc. Revascularization Med. 16 (3), 172–178.

Prati, F., Capodanno, D., Pawlowski, T., Ramazzotti, V., Albertucci, M., La Manna, A., Di Salvo, M., Gil, R.J., Tamburino, C., 2010. Local delivery versus intracoronary infusion of abciximab in patients with acute coronary syndromes. JACC: Cardiovasc. Interv. 3 (9), 928–934.

Remmelzwaal, L.A., Mishra, A.K., Ellis, G.F., 2020. Human eye inspired log-polar pre-processing for neural networks. In: 2020 International SAUPEC/RobMech/PRASA Conference. IEEE, pp. 1–6.

Sifre, L., Mallat, S., 2013. Rotation, scaling and deformation invariant scattering for texture discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1233–1240.

Souteyrand, G., Arbustini, E., Motreff, P., Gatto, L., Di Vito, L., Marco, V., Amabile, N., Chisari, A., Kodama, T., Romagnoli, E., et al., 2015. Serial optical coherence tomography imaging of ACS-causing culprit plaques. EuroIntervention 11 (3), 319–324.

Sun, K., Zhang, J., Liu, J., Yu, R., Song, Z., 2020. DRCNN: Dynamic routing convolutional neural network for multi-view 3D object recognition. IEEE Trans. Image Process. 30, 868–877.

Ughi, G.J., Adriaenssens, T., Sinnaeve, P., Desmet, W., D'hooge, J., 2013. Automated tissue characterization of in vivo atherosclerotic plaques by intravascular optical coherence tomography images. Biomed. Opt. Express 4 (7), 1014–1030.

Vergallo, R., Porto, I., De Maria, G.L., D'Amario, D., Annibali, G., Galli, M., Migliaro, S., Buccimazza, G., Aurigemma, C., Leone, A.M., et al., 2019. Dual quantitative coronary angiography accurately quantifies intracoronary thrombotic burden in patients with acute coronary syndrome: comparison with optical coherence tomography imaging. Int. J. Cardiol. 292, 25–31.

Wei, J., Xia, Y., Zhang, Y., 2019. M3net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. Pattern Recognit. 91, 366–378.

Xu, C., Zhao, W., Zhao, J., Guan, Z., Song, X., Li, J., 2022. Uncertainty-aware multiview deep learning for internet of things applications. IEEE Trans. Ind. Inform. 19 (2), 1456–1466.

Yan, C., Gong, B., Wei, Y., Gao, Y., 2020. Deep multi-view enhancement hashing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 43 (4), 1445–1451.

Yang, X., Feng, S., Wang, D., Zhang, Y., 2020. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans. Multimed. 23, 4014–4026.

Zhou, Y., Ye, Q., Qiu, Q., Jiao, J., 2017. Oriented response networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 519–528.

Zhou, J., Yu, S., Zhou, P., Liu, C., Sheng, Z., Li, J., Chen, R., Yan, H., Zhao, S., 2021. Impact of residual thrombus burden on ventricular deformation after acute myocardial infarction: A sub-analysis from an intravascular optical coherence tomography study. EClinicalMedicine 39.