Multi-view Contour-constrained Transformer Network for Thin-cap Fibroatheroma Identification

Liu, Sijie; Xin, Jingmin; Wu, Jiayi; Deng, Yangyang; Su, Ruisheng; Niessen, Wiro J.; Zheng, Nanning; van Walsum, Theo

DOI
10.1016/j.neucom.2022.12.041

Publication date
2023

Document Version
Final published version

Published in
Neurocomputing

Citation (APA)
Liu, S., Xin, J., Wu, J., Deng, Y., Su, R., Niessen, W. J., Zheng, N., & van Walsum, T. (2023). Multi-view Contour-constrained Transformer Network for Thin-cap Fibroatheroma Identification. *Neurocomputing*, *523*, 224-234. https://doi.org/10.1016/j.neucom.2022.12.041

Neurocomputing 523 (2023) 224–234

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-view Contour-constrained Transformer Network for Thin-cap Fibroatheroma Identification

Sijie Liu [a,c], Jingmin Xin [a,*], Jiayi Wu [a], Yangyang Deng [b], Ruisheng Su [c], Wiro J. Niessen [c,d], Nanning Zheng [a], Theo van Walsum [c]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
[b] Cardiovascular Department, First Affiliated Hospital of Medical College, Xi'an Jiaotong University, Xi'an 710049, China
[c] Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
[d] Faculty of Applied Sciences, Delft University of Technology, The Netherlands

ARTICLE INFO

Article history:
Received 7 April 2022
Revised 17 October 2022
Accepted 18 December 2022
Available online 23 December 2022
Communicated by Zidong Wang

Keywords:
IVOCT
TCFA
Plaque identification
Multi-view learning
Transformer

ABSTRACT

Identification and detection of thin-cap fibroatheroma (TCFA) from intravascular optical coherence tomography (IVOCT) images is critical for treatment of coronary heart diseases. Recently, deep learning methods have shown promising successes in TCFA identification. However, most methods usually do not effectively utilize multi-view information or incorporate prior domain knowledge. In this paper, we propose a multi-view contour-constrained transformer network (MVCTN) for TCFA identification in IVOCT images. Inspired by the diagnosis process of cardiologists, we use contour constrained self-attention modules (CCSM) to emphasize features corresponding to salient regions (i.e., vessel walls) in an unsupervised manner and enhance the visual interpretability based on class activation mapping (CAM). Moreover, we exploit transformer modules (TM) to build global-range relations between two views (i.e., polar and Cartesian views) to effectively fuse features at multiple feature scales. Experimental results on a semi-public dataset and an in-house dataset demonstrate that the proposed MVCTN outperforms other single-view and multi-view methods. Lastly, the proposed MVCTN can also provide meaningful visualization for cardiologists via CAM.

1. Introduction

Thin-cap fibroatheroma (TCFA) is defined as a lipid plaque with a fibrous cap less than $65\mu m$ thickness that forms in vessel walls and destroys the three-layer membrane structure of the vessel wall [1]. TCFA is also one type of vulnerable plaque that can lead to thrombosis, acute coronary syndrome, acute myocardial infarction or even sudden death [2]. Identification and detection of TCFA is critical for treatment of coronary heart diseases. Intravascular optical coherence tomography (IVOCT) is a catheter-based inspection method that uses near-infrared light to obtain high-resolution imaging of microstructures of blood vessel walls [3]. Compared with intravascular ultrasound (IVUS) imaging, IVOCT can acquire higher resolution images with clearly visible fine structures and characteristics of vulnerable plaques, so it has been widely used as a gold standard for assessing TCFA [4]. Usually, a catheter that contains an OCT probe is inserted into a coronary artery and then pulled back, obtaining hundreds of high-resolution images. Manual TCFA identification from so many images is not only time-consuming but also subjective. Therefore, automated TCFA identification in IVOCT images is a valuable and challenging task.

Recently, deep learning methods have shown promising success in TCFA identification. Wang et al. [5] used a convolutional neural network to extract multi-scale features from a polar view of IVOCT image to identify vulnerable plaques. Xu et al. [6] studied the effectiveness of four types of deep neural networks in identifying fibroatheromas, where deep learning features were directly extracted from a Cartesian view of IVOCT image. Multi-view learning has recently received increasing attention. Gessert et al. [7] fused deep learning features from two views of images (i.e., the polar and Cartesian views of IVOCT images) via a single concatenation operation to improve the identification performance. However, there is huge spatial distortion between the two views, shown in Fig. 1. The same location in the two views of images may have completely unrelated clinical biomarkers. Fusing the features of the biomarkers through a concatenation operation may lead to suboptimization. Therefore, more effective fusion

* Corresponding author.
E-mail address: jxin@mail.xjtu.edu.cn (J. Xin).

https://doi.org/10.1016/j.neucom.2022.12.041
0925-2312/

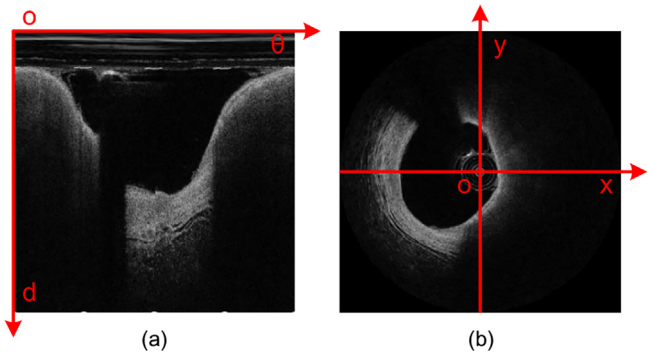

Fig. 1. A schematic diagram of the view transformation between (a) a polar view of image and (b) a Cartesian view of image.

operations may be considered and multi-scale feature fusion may also be an effective factor for TCFA identification.

In addition, there have been some attempts to incorporate prior domain knowledge into deep learning. Shi et al. [8] exploited two cascaded networks to identify vulnerable plaques according to a diagnosis process of cardiologists in which cardiologists first focus on salient regions and then locate and identify vulnerable plaques in the salient regions. Liu et al. [9] proposed to unify the two networks as a single network through sharing the bottom layers of the two networks. However, both methods require additional salient regions to be annotated for the training phase, which increases the annotation burdens on cardiologists and limits the application of the methods. Therefore, a method without annotating the salient region deserves further study.

In this paper, we consider incorporating prior domain knowledge in an unsupervised manner and fusing the multi-view features at multiple feature scales for TCFA identification. To this end, we propose a multi-view contour-constrained transformer network (MVCTN) for TCFA identification in IVOCT images. First, inspired by the diagnosis process of cardiologists, we use contour constrained self-attention modules (CCSM) to emphasize features corresponding to salient regions (i.e., vessel walls) in an unsupervised manner and enhance the visual interpretability based on class activation mapping (CAM) [10]. Furthermore, we use transformer modules (TM) to build global-range relations between the two views (i.e., the polar view and the Cartesian view) to effectively fuse the two views of features at multiple feature scales. Specifically, as shown in Fig. 2, we first use a view transformation to generate the two views of images. Second, the polar view is used as a primary view while the Cartesian view is used as an auxiliary view (or vice versa). Third, each of the views has a SENet-based pathway with three CCSMs, each of which includes a spatial attention block (SAB) and an active contour loss (ACL). The ACL can support the SAB to improve the detection ability of the salient regions as well as the visual interpretability based on CAM. It's worth noting that since TCFA forms in vessel walls and destroys normal vessel wall structures in IVOCT images, the vessel wall is considered as the salient region in this paper. Finally, three TMs bridge the two pathways to allow the auxiliary view to assist the primary view at the multiple feature scales. In order to verify effectiveness and robustness of the proposed MVCTN, we evaluate it on one semi-public dataset (i.e., the 2017 Chinese Conference on Computer Vision-IntraVascular Optical Coherence Tomography challenge (CCCV-IVOCT) dataset) and one in-house dataset (i.e., the Optical coherence tomography Plaque Recognition Database (OPRD)). Besides, we also try to use CAM to understand and interpret decisions of the proposed MVCTN. In summary, main contributions are as follows:

- We propose a multi-view contour-constrained transformer network (MVCTN), which emphasizes the features corresponding to the salient regions in the unsupervised manner, enhances the visual interpretability based on CAM and fuses the multi-view features at the multiple feature scales for TCFA identification.

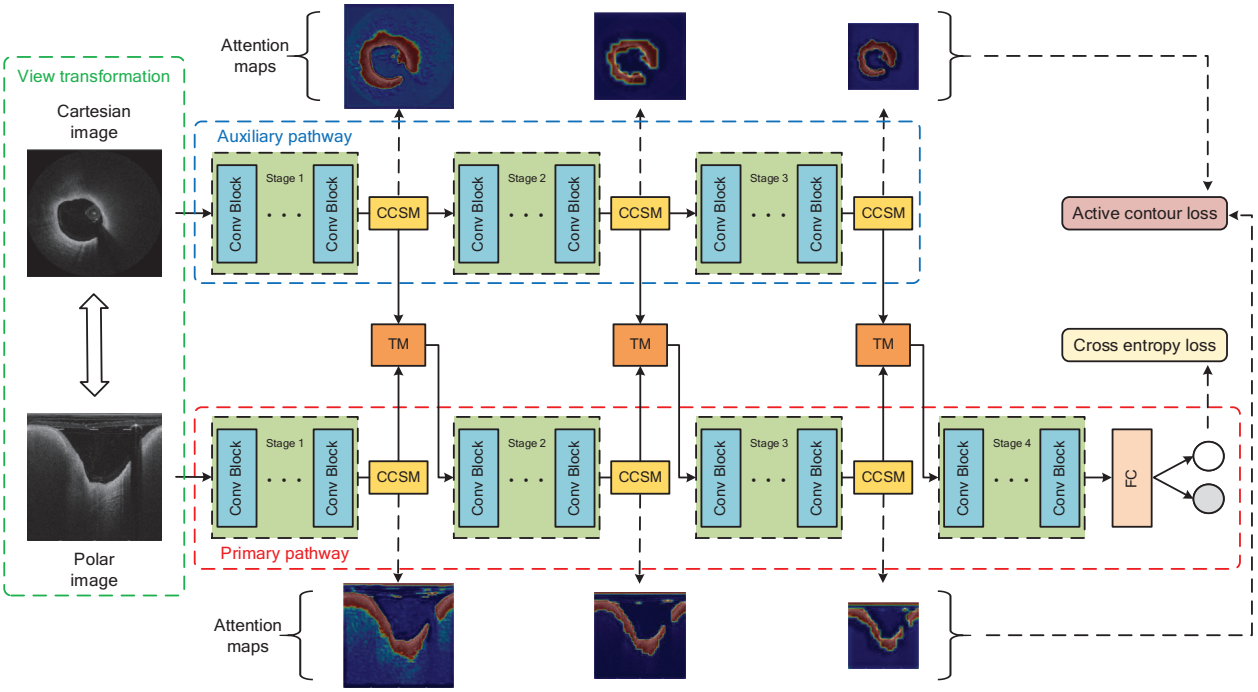

Fig. 2. An overview of the proposed method for TCFA identification in the IVOCT images. In this overview, we set the polar view as the primary view, and the Cartesian view as the auxiliary view. **In fact, the primary view and the auxiliary view are interchangeable.**

- Inspired by the diagnosis process of cardiologists, we propose a CCSM consisting of a SAB and an ACL. The SAB is a type of self-attention block which can emphasize or suppress features in different spatial locations by learning. The ACL is attached to the SAB to improve the detection ability of the salient regions and enhance the visual interpretability based on CAM.
- In order to effectively fuse the two views of features with huge spatial distortion, we introduce three TMs to build the global-range relation between the two views at the multiple feature scales.
- The proposed MVCTN not only outperforms other single-view and multi-view methods on the two datasets, but also provides meaningful CAM-based visualization, thus potentially becoming a useful tool with explainable results for cardiologists, especially for inexperienced cardiologists, to support them in making diagnostic decisions.

2. Related Work

2.1. Plaque Identification in IVOCT Images

Automated methods for plaque identification in IVOCT images can be generally divided into three categories: **polar view based methods, Cartesian view based methods and multi-view based methods.** The polar view based methods only use the polar view of images. For example, Rico-Jimenez et al. [12] segmented lumen areas from the polar view of images, and classified plaque types based on their morphological characteristics. With the rapid development of deep learning, Liu et al. [13] proposed a ResNet-3D network to classify coronary plaques in the polar view of IVOCT pullbacks. And Shi et al. [14] proposed a deep multiple instance learning method to classify and locate vulnerable plaques from the polar view of IVOCT images. The Cartesian view based methods only use the Cartesian view of images. For example, texture features were directly extracted from the Cartesian view of images to automatically identify plaques [15]. Recently, convolutional neural networks (CNNs) were proposed to automatically learn plaque-related features from the Cartesian view of images [16,17]. The multi-view based methods use both of the polar and Cartesian views of images. For example, Xu et al. [18] transformed the Cartesian view of images into the polar view of images and extracted texture features from the polar view of images to identify plaques. In contrast, Zhou et al. [19] segmented lumen contours using the polar view of images and transformed the segmented lumen contours into the Cartesian representation to classify plaque tissues. Furthermore, deep multi-view learning for the plaque identification has drawn some attention. Gessert et al. [7] demonstrated that the performance of plaque identification could be improved by fusing deep learning features from both of the polar and Cartesian views of images.

2.2. Active Contour Model

Active contour models, also known as snakes [20], evolve a contour to detect objects in a given image by minimizing an energy function based on marginal information and smoothness constraints. However, snakes are sensitive to initial contour location and might get stuck in a local minimum. In the past three decades, a number of snake variants have been proposed, such as active contour without edges (ACWE) [21], where an implicit contour evolution modelled as the evolution of a zero level set is guided by the optimization of an energy function. The active contour model has been widely used in medical image segmentation [22]. Some of the methods based on deep learning used the active contour model as an active contour loss (ACL) to optimize deep neural networks [23,24]. Inspired by this, we apply an ACL to identify TCFA in IVOCT images.

2.3. Transformer

The architecture of Transformer [25] can effectively capture long-range dependencies and has become a de facto standard for natural language processing tasks. There are also many variants of Transformer in the computer vision community. Wang et al. [26] proposed a non-local module to enhance CNN's ability to capture long-range dependencies. Dosovitskiy et al. [27] replaced CNNs with the architecture of Transformer. Inspired by this, in this paper, we propose a new variant of Transformer to build global-range relations between the polar view and the Cartesian view, thereby overcoming the huge spatial distortion between the two views.

3. Methods

An overview of the proposed method is illustrated in Fig. 2. Firstly, we use a view transformation to generate the two views of images. Secondly, we insert CCSMs into a SENet [28] to build two types of SENet-based pathways, including a primary pathway and an auxiliary pathway. The primary pathway is a SENet where three CCSMs are inserted behind the $1^{st}, 2^{nd}$, and 3^{rd} convolution stages, respectively. The auxiliary pathway has a similar structure to the primary pathway, and their difference is that there is no module after the 3rd CCSM of the auxiliary pathway. Thirdly, we use three TMs to bridge the two pathways to allow an auxiliary view to assist a primary view at the multiple feature scales. It's worth noting that the primary view and the auxiliary view are interchangeable. Finally, we propose a loss function including one cross entropy loss and six active contour losses for optimizing the proposed MVCTN.

3.1. View Transformation

In order to utilize multi-view information, we use a view transformation to generate the other view of images from one view of images. For the polar view of images, the Cartesian transformation [7] is applied. Specifically, the polar image $I_p(d,\theta)$ can be transformed into Cartesian representation with the transformation $x\prime = dcos(\theta)$ and $y\prime = dsin(\theta)$. Then the Cartesian image $I_c(x,y)$ is derived from applying the bilinear interpolation in the transformed image $I_c(x\prime,y\prime)$. For the Cartesian view of images, the polar transformation is applied. Specifically, the polar image $I_p(d,\theta)$ is derived from the transformation $d = x/cos(arctan(y/x))$ and $\theta = arctan(y/x)$. The schematic diagram of the view transformation between the two views is illustrated in Fig. 1. The polar view of images do not suffer from distortion caused by the bilinear interpolation artifacts, while using the Cartesian view of images is more intuitive as their representation resembles the anatomical structure of the artery. We hypothesize that one view of representation can be a useful complement to the other view of representation in deep learning methods.

3.2. Contour Constrained Self-attention Module

A contour constrained self-attention module (CCSM) can emphasize features corresponding to the salient regions (i.e. the vessel walls) in an unsupervised manner and enhance the visual interpretability based on CAM. The CCSM consists of a spatial attention block (SAB) and an active contour loss (ACL), shown in Fig. 3. The SAB is a type of self-attention block which can emphasize or suppress features in different spatial locations by learning.

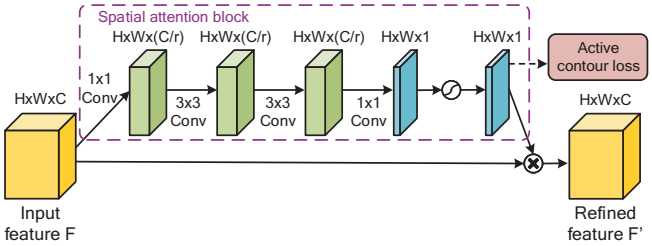

Fig. 3. A schematic diagram of the contour constrained self-attention module. Sizes of features are also indicated.

The ACL is attached to the SAB to improve the detection ability of the salient regions and enhance the visual interpretability based on CAM. The following describes the details of the SAB and ACL.

3.2.1. Spatial Attention Block

Given an input feature $F \in \mathbb{R}^{H\times W\times C}$, the SAB infers an attention map $S(F) \in [0,1]^{H\times W\times 1}$. The refined feature $F\prime$ is formulated as:

$$F\prime = F \otimes S(F), \tag{1}$$

where $\otimes$ denotes element-wise multiplication while S indicates four convolution layers and one sigmoid fuction.

3.2.2. Active Contour Loss

Compared with ACWE [21], the ACL removes the two regularization terms with respect to the length and area as [11] does, and also adds a constraint to ensure the foreground is the salient region. Given a shifted attention map $\phi = S(F) - 0.5 \in [-0.5, 0.5]^{H\times W\times 1}$, the ACL infers a loss value $ACL(\phi) \in \mathbb{R}$ which is formulated as:

$$ACL(\phi) = \int_\Omega |u(x,y) - c_1|^2 H_\epsilon^*(\phi(x,y))dxdy + \int_\Omega |u(x,y) - c_2|^2 (1 - H_\epsilon^*(\phi(x,y)))dxdy,$$
$$s.t. \quad c_1 \geqslant c_2, \tag{2}$$

where Ω denoted an entire domain of a given image u that has been downsampled to the same size as the shifted attention map ϕ, and $u(x,y)$ was a pixel value at a location $(x,y) \in \Omega$. And c_1 and c_2 are the average pixel values inside and outside the contour, respectively, which are computed as:

$$\begin{cases} c_1(\phi) = \dfrac{\int_\Omega u(x,y) H_\epsilon^*(\phi(x,y))dxdy}{\int_\Omega H_\epsilon^*(\phi(x,y))dxdy}, \\ c_2(\phi) = \dfrac{\int_\Omega u(x,y)(1 - H_\epsilon^*(\phi(x,y)))dxdy}{\int_\Omega (1 - H_\epsilon^*(\phi(x,y)))dxdy}, \end{cases} \tag{3}$$

where the modified approximated Heaviside function (MAHF) [11] H_ϵ^* can allow that errors can be back propagated to previous layers, which is defined as:

$$H_\epsilon^*(z) = \frac{1}{2}(1 + \tanh(\frac{z}{\epsilon})). \tag{4}$$

It's worth noting that since the SAB isn't directly supervised by any annotation information of the salient regions, a totally inverse shifted attention map ϕ, focusing on the non-salient regions, may be generated by the SAB, resulting in $c_1(\phi) < c_2(\phi)$. As we can observe from Fig. 4, the vessel walls are surrounded exactly by the red contours but the corresponding shifted attention maps show that the other areas should be emphasized. Hence, if this issue occurs, the attention map $S(F)$ will be replaced by $1 - S(F)$, causing ϕ to be transformed to $-\phi$. In this way, the constraint $c_1 \geqslant c_2$ in Eq. (2) can always be satisfied, and each attention map of the SAB with the ACL is shown in Fig. 2.

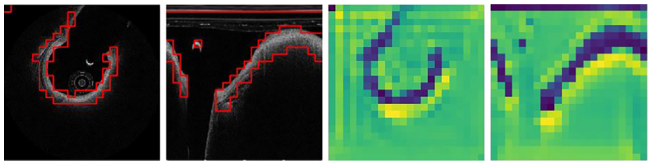

Fig. 4. Visualized outputs of the CCSM without the constraint $c_1 \geqslant c_2$. The left two images show the red contours generated by the CCSM while the right two images show the corresponding shifted attention maps. The yellow, green and blue pixels denote the high, medium, low probabilities, respectively.

It is easy to optimize the ACL defined in Eq. (2). By calculus of variation, the derivative of the ACL with respect to ϕ can be written as:

$$\frac{\partial ACL}{\partial \phi} = \delta_\epsilon^*(\phi)[(u_0 - c_1)^2 - (u_0 - c_2)^2], \delta_\epsilon^*(z) = \frac{\partial H_\epsilon^*(z)}{\partial z} = \frac{1}{2\epsilon}\left(1 - \tanh\left(\frac{z}{\epsilon}\right)\right)\left(1 + \tanh\left(\frac{z}{\epsilon}\right)\right). \tag{5}$$

3.3. Transformer Module

A transformer module (TM) builds the global-range relation between the two views to allow the auxiliary view to assist the primary view at the feature scale, shown in Fig. 5. Compared to the naive concatenation operation in [7], the TM overcomes the huge spatial distortion between the two views by calculating the global-range relation, and thus does not distort the spatial distribution of the primary view of features, which will be demonstrated in Section 5.1.

Specifically, given a primary view of feature $x_p \in \mathbb{R}^{H\times W\times C}$ and an auxiliary view of feature $x_a \in \mathbb{R}^{H\times W\times C}$, x_p is transformed into a

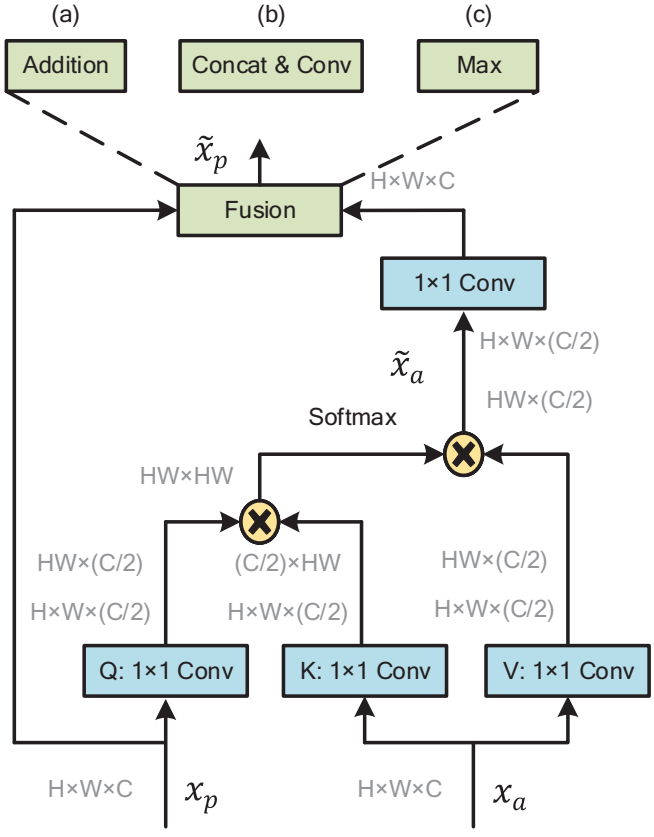

Fig. 5. An architecture of the transformer module. There are three alternative versions of fusion operations, including (a) addition, (b) concatenation followed by convolution, and (c) max.

query matrix $Q \in \mathbb{R}^{H\times W\times(C/2)}$, while x_a is transformed into a key matrix $K \in \mathbb{R}^{H\times W\times(C/2)}$ and a value matrix $V \in \mathbb{R}^{H\times W\times(C/2)}$. Then a weight matrix applied to the matrix V from x_a is determined by the matrix multiplication between the matrix Q from x_p and the matrix K from x_a, which can capture the relations between each position in x_p and all the positions in x_a. And thus the weighted feature $\tilde{x}_a$ is calculated as:

$$\tilde{x}_a = \mathrm{Softmax}(\mathrm{QK}^{\mathrm{T}})\mathrm{V}. \tag{6}$$

Further, the weighted feature $\tilde{x}_a$ undergoes one convolution to produce the same dimensions as x_p, and then is fused with the original primary view of feature x_p to output the fused feature $\tilde{x}_p$ that is defined as:

$$\tilde{x}_p = \mathrm{Fuse}(\mathrm{C}(\tilde{\mathrm{x}}_{\mathrm{a}}),\ \mathrm{x}_{\mathrm{p}}), \tag{7}$$

where C denotes a convolution operation, while Fuse denotes a fusion operation. Herein, we provide three alternative versions for the fusion operation, including (a) addition, (b) concatenation followed by convolution, and (c) max [29].

3.4. Loss function

The proposed MVCTN can be optimized by one cross entropy (CE) loss and six active contour losses (ACL). The total loss function is formulated as:

$$Loss = CE(\tilde{y}_p, y) + \lambda\left(\sum_{n=1}^{3} ACL(\phi_{u_p}^n) + ACL(\phi_{u_a}^n)\right), \tag{8}$$

where subscripts p and a indicate the primary view and auxiliary view respectively, $\tilde{y}$ and y denote a predicted probability and its corresponding label respectively, ϕ_u^n denotes a shifted attention map generated by the n^{th} CCSM with an image u as input, and λ is a loss coefficient.

4. Experiments

4.1. Dataset Description

4.1.1. CCCV-IVOCT

The 2017 Chinese Conference on Computer Vision-IntraVascular Optical Coherence Tomography challenge (CCCV-IVOCT) dataset is a semi-public dataset for detecting TCFA in IVOCT images, whose data is supplied by Xi'an Institute of Optics and Precision Mechanics of Chinese Academy of Sciences. This dataset is not in the public domain, but the institute grants participants of this challenge the right to use this dataset and publish results of this dataset. To the best of our knowledge, many studies on TCFA segmentation or detection [8,30,31] have been conducted on this dataset. But in this paper, we consider the problem of identifying whether an IVOCT image contains TCFA.

This dataset consists of 2300 IVOCT images represented in the polar coordinate system with a size of 352×720 pixels. Specifically, there are 1000 images with TCFA (i.e., positive samples) and 1000 images without TCFA (i.e., negative samples) in a training set, while there are 198 positive samples and 102 negative samples in an independent testing set. Given the relatively small size of this dataset, we use a 5-fold cross-validation strategy to evaluate the proposed MVCTN on this dataset. Specificaly, in every fold, the training set is divided into a new training set with 1800 images and a validation set with 200 images. The division of this dataset is summarized in Table 1. Then the model that performs best on the validation set is further evaluated on the testing set. Finally, the mean and standard deviation performance of the five folds are reported.

Table 1
Division of the CCCV-IVOCT dataset and OPRD. P/N: positives/Negatives.

Dataset	Training (P/N)	Validation (P/N)	Testing (P/N)	View
CCCV-IVOCT	900/900	100/100	198/102	Polar
OPRD	1946/2054	590/710	663/537	Cartesian

4.1.2. OPRD

The Optical coherence tomography Plaque Recognition Database (OPRD) is our in-house dataset and includes 31049 IVOCT images from 2137 vessel segments of 540 patients supplied by the Second Affiliated Hospital of Harbin Medical University. The patients are between 18 to 80 years old and have not undergone stent surgery. These IVOCT images are obtained through OPTIS from June 2015 to August 2016, and are represented in the Cartesian coordinate system. TCFA identification from the IVOCT images is performed by two MD students and one chief physician with more than 10 years of clinical experience, who use the C7-XR/ILUMIEN OCT system to detect fibrous caps less than $65\mu m$ thickness according to the guideline of Sinclair et al. [1]. The two MD students identify TCFA independently and the chief physician independently identifies cases in which there is disagreement by the two MD students.

Since images within a single vessel segment have a high degree of similarity and there is a huge imbalance between the positive and negative samples, we sample 6500 images from 31049 images. These sampled images are divided into a training set with 4000 images, a validation set with 1200 images, and a testing set with 1300 images. Each of the three sets is obtained by sampling from different vessel segments and thus is independent. The division of this dataset is summarized in Table 1.

4.2. Implement Details and Evaluation Metrics

4.2.1. Implement details

During the training phase, the polar and Cartesian views of images are generated by the view transformation and resized as $352 \times 352 \times 3$ pixels for initializing the pre-trained weights of ImageNet. The data augmentation scheme depends on the view. For the polar view, the scheme includes circular translation and horizontal reflection. For the Cartesian view, the scheme includes rotation, horizontal and vertical reflection. The proposed MVCTN is implemented with Pytorch [32], and is trained on two Nvidia GPUs (i.e., GTX Titan X) with a batch size of 20. The proposed MVCTN is initialized with the pre-trained weights of ImageNet, as [7] does. In addition, an adaptive moment estimation (Adam) with an initial learning rate of 10^{-3}, a momentum of 0.9 and a weight decay of 5×10^{-4} is used for optimization. Lastly, the loss coefficient λ is set to 2. During the inference phase, no data augmentation scheme or post-processing operation is applied.

4.2.2. Evaluation Metrics

According to the characteristics of the datasets, we use different metrics to evaluate the proposed MVCTN on the two datasets. For the CCCV-IVOCT dataset, since there is an imbalanced ratio between the positive and negative samples (about 2:1) in the testing set, we use the area under the receiver operating characteristic curve (AUC) as a main metric, and also show results on F1-score, precision, and recall. For OPRD that is more balanced, we use accuracy, F1-score, precision, and recall as the evaluation metrics.

4.3. Ablation Studies

In this subsection, an ablation study is conducted on the CCCV-IVOCT dataset to investigate the effectiveness of our design

choices. We first study the impact of the loss coefficient of the ACL of the CCSM, then explore the influences of the quantity and the fusion operation of the TM. Finally, we study the effects of the CCSM and TM in a quantitative way.

4.3.1. Loss Coefficient of ACL

We study the impact of the loss coefficient λ of the ACL on the single-view network (i.e., SENet50 + CCSM). Fig. 6 shows the results of the polar and Cartesian views with six different loss coefficients λ (i.e., 0, 0.5, 1, 2, 3, 5). When $\lambda = 0$, the single-view network is only trained with the cross entropy loss, and the SAB is optimized by the gradient from the cross entropy loss. When $\lambda > 0$, the single-view network is trained with both of the cross entropy loss and the active contour loss (ACL), and the SAB is optimized by the two losses. As shown in Fig. 6, the AUCs of both views drop when λ is small, next rise gradually, then reach the maximum values (0.8904 for the polar view and 0.8990 for the Cartesian view) at $\lambda = 2$, finally decrease by a small margin. The result suggests that it's relevant to determine a suitable loss coefficient λ. λ is set to 2 in the follow-up experiments unless otherwise specified.

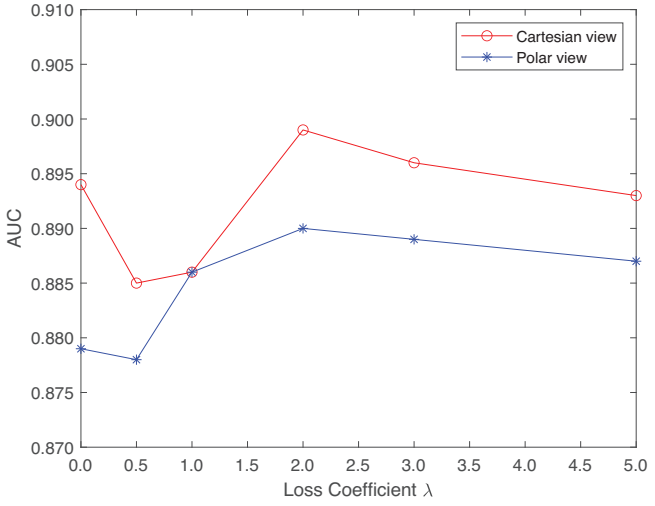

Fig. 6. Results of the polar and Cartesian views with different loss coefficients λ.

4.3.2. Quantity and Fusion Operation of TM

Based on the above study, we further explore the influences of the quantity and the fusion operation of the TM on the proposed MVCTN in the Cartesian view. We bridge three TMs between the three pairs of CCSMs or bridge one TM between a pair of CCSMs inserted behind the 3^{th} convolution stage of SENet50. And we provide three alternative fusion operations for the TM, including Addition, Concat & Conv, and Max. Table 2 shows the results with the different fusion operations and the different number of TMs. It's observed that the more TMs lead to the better AUC, F1-score and recall in general, and that Max obtains the better average AUC and F1-score compared to Addition and Concat & Conv. As a result, three TMs with Max can lead to the best average AUC (91.31%), F1-score (90.76%), and recall (97.98%). Thus Max is used as the fusion operation of the TM in the follow-up experiments unless otherwise specified.

4.3.3. Effects of CCSM and TM

We further study the effects of the CCSM and TM in a quantitative way. We divide the experiments into two groups: one takes the polar view as the primary view and the other takes the Cartesian view. The results are reported in Table 3. Firstly, compared with SENet50, SENet50 armed with the CCSM boosts the results of all the evaluation metrics in both groups. Then compared with SENet50, SENet armed with the TM enhances the results of all evaluation metrics in both groups as well. Finally, SENet50 armed with both of the CCSM and TM achieves the best AUC, F1-score and recall in both groups.

4.4. Comparisons with the other methods

Single-view methods: To investigate the effectiveness of the CCSM, we compare with EfficientNet-B0 [33], ResNet50 [34], SENet50 [28], and Twins [38].

Multi-view methods: According to the taxonomy [35], existing methods for multimodal fusion can be categorized into three

Table 2
Results of the Cartesian view with the different fusion operations and the different number of TMs.

Operation	Quantity	AUC(%)	F1-score(%)	Precision(%)	Recall(%)
Addition	1	90.85±0.57	89.89±2.14	87.81±3.30	92.63±4.64
Concat & Conv	1	90.89±0.49	88.98±2.29	86.55±5.11	92.32±5.12
Max	1	91.13±0.37	90.48±2.57	**89.30±1.65**	92.42±3.64
Addition	3	91.03±0.56	90.09±1.76	87.01±3.32	93.84±3.65
Concat & Conv	3	91.15±0.45	90.55±1.08	86.76±3.68	95.55±3.74
Max	3	**91.31±0.35**	**90.76±0.78**	84.71±3.68	**97.98±3.29**

Table 3
Ablation results on the CCCV-IVOCT dataset. The best results are in bold.

SENet50	CCSM	TM	View		AUC(%)	F1-score(%)	Precision(%)	Recall(%)
			Primary	Auxiliary				
√			Polar	-	87.43±1.85	84.11±4.04	89.06±2.04	80.51±7.87
√	√		Polar	-	89.04±0.84	87.30±1.29	89.98±1.26	84.85±3.05
√		√	Polar	Cartesian	89.97±0.67	88.27±2.13	89.64±3.64	87.47±4.47
√	√	√	Polar	Cartesian	90.56±0.39	90.18±1.18	87.63±3.20	93.33±1.57
√			Cartesian	-	88.41±1.61	83.41±4.82	89.05±1.48	79.70±6.50
√	√		Cartesian	-	89.90±0.40	87.26±2.52	90.24±2.21	84.95±4.13
√		√	Cartesian	Polar	90.41±0.41	88.11±3.83	**90.52±1.94**	86.16±5.01
√	√	√	Cartesian	Polar	**91.31±0.35**	**90.76±0.78**	84.71±3.68	**97.98±3.29**

classes: input-level fusion, feature-level fusion, and decision-level fusion. Therefore, to investigate the effectiveness of the TM, we compare with a straightforward input-level fusion method (SIFM) (Fig. 7(a)), a straightforward feature-level fusion method [7] (SFFM) (Fig. 7a straightforward decision-level fusion method (SDFM) (Fig. 7h are based on SENet50. Besides, we also compare with one advanced input-level fusion method (i.e., TransMed [36]) and one advanced feature-level fusion method (i.e., BFNet [37]).

4.4.1. Results on the CCCV-IVOCT Dataset

Experimental results are reported in Table 4. The proposed single-view network is SENet50 + CCSM and the proposed multi-view networks are the MVCTN (p) and the MVCTN (c). The MVCTN (p) takes the polar view as the primary view while the MVCTN (c) takes the Cartesian view as the primary view. First of all, SENet50 + CCSM outperforms the other single-view methods on AUC, F1-score, and recall, and obtains the comparable result on precision. We argue that the CCSM can force SENet50 to focus on the features corresponding to the salient regions (i.e., the vessel walls), and thus improves the performance. Moreover, from the comparison among the multi-view methods, we can find that the MVCTN (c) achieves the best AUC (91.31%), F1-score (90.76%), and recall (97.98%), and that the MVCTN (p) also yields the second best AUC (90.56%), F1-score (90.18%), and recall (93.33%). This demonstrates that using TM to fuse the two views of features is an appropriate choice in this task.

4.4.2. Results on OPRD

To further assess the effectiveness and robustness of the proposed MVCTN, we also compare with the aforementioned methods on OPRD. Table 5 shows these experimental results. Firstly, SENet50 + CCSM outperforms the other single-view methods on AUC, F1-score, and recall, and obtains the comparable result on precision. Secondly, the MVCTN (p) achieves the best accuracy (93.51%) and F1-score (93.68%), and the MVCTN (c) also yields the second best accuracy (93.20%) and F1-score (93.40%). These results show the proposed networks have good effectiveness and robustness when applied to a new dataset.

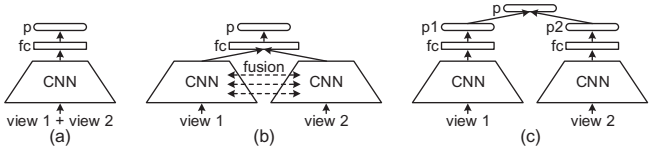

Fig. 7. Three types of multi-view fusion methods: (a) an input-level fusion method, (b) a feature-level fusion method, and (c) a decision-level fusion method.

5. Discussion

5.1. Visualization based on CAM

We analyze the effects of the CCSM and TM on the visualization based on CAM for correctly classified samples, which is shown in Fig. 8. For negative samples, we visualize negative class activation maps (CAMs). For positive samples, we visualize positive CAMs. In every case, the Cartesian and polar views of image can be transformed mutually. For SENet50 + TM and SENet50 + CCSM + TM, the CAM is generated by taking the input image in the same column as the primary view of image, and the other input image as the auxiliary view of image. For SENet50 and SENet50 + CCSM, the CAM is generated from the input image in the same column.

Firstly, we find that for the negative samples, SENet50 presents the CAMs where most of the regions, including some regions irrelevant to the TCFA identification in clinical diagnosis, have the high heat values. However, SENet50 + CCSM and SENet50 + CCSM + TM have the high heat values on the regions with the normal vessel wall structure. Secondly, for the positive samples, the CAMs of all the methods show that the region of TCFA is highlighted while the other regions are not, indicating a good ability of detecting TCFA. However, compared to SENet50, SENet50 + CCSM and SENet50 + CCSM + TM tend to ignore catheter regions with high pixel values. Thirdly, the CAMs of SENet50 + TM are similar to the CAMs of SENet50, demonstrating that a key to improving the visualization result is CCSM instead of TM. Last but not least, CAMs generated by SENet50 + TM and SENet50 + CCSM + TM are based on the primary view designated by users. It manifests that using the TM to fuse the two views of features doesn't distort the spatial distribution of the primary view of features. The comparison between SFFM [7] and SENet50 + TM in Fig. 9 further demonstrates this conclusion. SFFM, which uses a concatenation operation to fuse the two views of features, produces the high heat values in regions that overlap with the regions with the normal vessel wall structure of the one of the two views. But SENet50 + TM (p) produces the high heat values only in regions that overlap with those of the polar view, and SENet50 + TM (c) produces the high heat val-

Table 4
Comparison with the other methods on the CCCV-IVOCT dataset. The best results are in bold.

Method	View	Fusion	AUC(%)	F1-score(%)	Precision(%)	Recall(%)
EfficientNet-B0 [33]	Polar	-	87.17±3.07	77.54±5.52	90.91±1.50	68.08±7.38
ResNet50 [34]	Polar	-	87.51±1.89	76.32±3.61	89.95±1.15	67.58±8.42
SENet50 [28]	Polar	-	87.43±1.85	84.11±4.04	89.06±2.04	80.51±7.87
Twins [38]	Polar	-	88.21±1.21	86.13±2.82	89.75±1.38	82.62±4.50
SENet50 + CCSM	Polar	-	89.04±0.84	87.30±1.29	89.98±1.26	84.85±3.05
EfficientNet-B0 [33]	Cartesian	-	88.19±2.16	79.69±4.08	90.60±1.56	72.39±5.71
ResNet50 [34]	Cartesian	-	88.07±2.09	79.26±5.76	**90.93±1.77**	70.71±7.90
SENet50 [28]	Cartesian	-	88.41±1.61	83.41±4.82	89.05±1.48	79.70±6.50
Twins [38]	Cartesian	-	89.32±1.21	86.53±3.13	89.85±1.63	83.25±4.82
SENet50 + CCSM	Cartesian	-	89.90±0.40	87.26±2.52	90.24±2.21	84.95±4.13
SIFM (Fig. 7(a))	Polar + Cartesian	Input-level	88.93±0.62	86.83±3.28	89.32±2.04	84.85±4.31
TransMed [36]	Polar + Cartesian	Input-level	89.46±0.27	87.77±3.24	88.63±4.04	87.78±4.99
SFFM [7] (Fig. 7(b))	Polar + Cartesian	Feature-level	89.90±0.20	89.21±1.84	88.11±3.47	90.81±3.91
BFNet [37]	Polar + Cartesian	Feature-level	89.95±1.18	86.65±2.48	88.20±5.47	86.06±4.55
SDFM (Fig. 7(c))	Polar + Cartesian	Decision-level	89.59±0.60	88.36±1.20	87.68±5.55	89.69±3.41
MVCTN (p)	Polar + Cartesian	Feature-level	90.56±0.39	90.18±1.18	87.63±3.20	93.33±1.57
MVCTN (c)	Polar + Cartesian	Feature-level	**91.31±0.35**	**90.76±0.78**	84.71±3.68	**97.98±3.29**

Table 5
Comparison with the other methods on OPRD. The best results are in bold.

Method	View	Fusion	Accuracy(%)	F1-score(%)	Precision(%)	Recall(%)
EfficientNet-B0 [33]	Polar	-	92.00±0.56	92.06±0.32	93.20±0.51	90.95±2.23
ResNet50 [34]	Polar	-	91.46±0.27	91.35±0.87	**94.52±0.47**	88.39±3.10
SENet50 [28]	Polar	-	91.82±0.54	92.12±0.63	91.64±1.14	92.61±1.57
Twins [38]	Polar	-	92.33±0.52	92.44±0.42	91.86±0.92	93.03±1.45
SENet50 + CCSM	Polar	-	92.85±0.49	93.02±0.50	92.54±0.89	93.51±1.28
EfficientNet-B0 [33]	Cartesian	-	91.31±0.31	91.38±0.52	92.44±1.07	90.35±2.69
ResNet50 [34]	Cartesian	-	90.00±1.24	89.95±1.17	92.23±2.33	87.78±3.65
SENet50 [28]	Cartesian	-	90.92±0.89	91.07±0.97	91.35±1.56	90.80±2.54
Twins [38]	Cartesian	-	91.71±0.65	91.73±0.88	92.03±1.49	91.51±2.31
SENet50 + CCSM	Cartesian	-	92.23±0.57	92.40±0.48	92.19±1.89	92.61±1.63
SIFM (Fig. 7(a))	Polar + Cartesian	Input-level	91.95±0.82	92.30±0.78	91.95±2.31	93.06±1.73
TransMed [36]	Polar + Cartesian	Input-level	92.23±0.37	92.51±0.66	90.96±2.54	94.12±1.25
SFFM [7] (Fig. 7(b))	Polar + Cartesian	Feature-level	92.69±0.48	92.95±0.56	91.52±1.67	94.42±1.42
BFNet [37]	Polar + Cartesian	Feature-level	92.46±1.04	92.62±0.57	92.48±1.25	92.76±1.46
SDFM (Fig. 7(c))	Polar + Cartesian	Decision-level	92.32±0.26	92.55±0.37	91.15±1.53	**94.72±1.02**
MVCTN (p)	Polar + Cartesian	Feature-level	**93.51±0.53**	**93.68±0.69**	92.86±1.70	94.12±1.38
MVCTN (c)	Polar + Cartesian	Feature-level	93.20±0.49	93.40±0.22	92.31±0.95	94.12±1.19

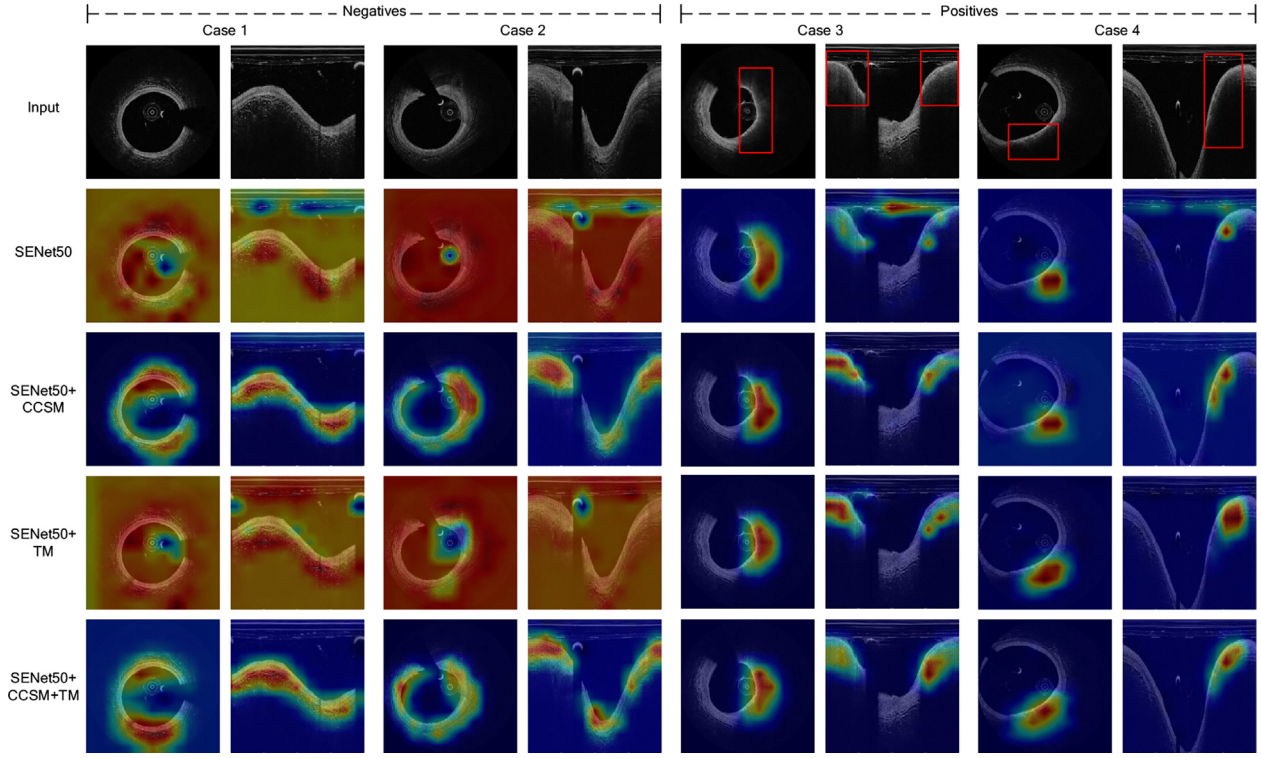

Fig. 8. CAM-based visualization results for correctly classified samples. In every case, the Cartesian view of image and the polar view of image can be transformed mutually. And in every case, for SENet50 + TM and SENet50 + CCSM + TM, the CAM is generated by taking the input image in the same column as the primary view of image, and the other input image as the auxiliary view of image. For SENet50 and SENet50 + CCSM, the CAM is generated from the input image in the same column. In addition, the regions of TCFA are indicated with the red box.

ues only in regions that overlap with those of the Cartesian view. This characteristic makes the CAMs of SENet50 + TM more explainable.

We also visualize and analyze some misclassified samples of the proposed MVCTN, shown in Fig. 10. For the negative samples, we visualize the positive CAMs, because these negative samples are misclassified as the positive samples. Similarly, for the positive samples, we visualize the negative CAMs. Firstly, the proposed MVCTN predicts the high positive probabilities (0.93 and 0.95) for the two misclassified negative samples, and their CAMs look similar to the CAMs of true positive samples, where a small region within the vessel wall is highlighted. According to our cardiologist's analysis, the two highlighted regions actually contain lipid plaques. The lipid plaques have some characteristics of TCFA but the thickness of their fibrous cap is greater than 65μm. We consider that these lipid plaques are potentially dangerous and also worthy of cardiologist's attention. Secondly, the CAMs of the two misclassified positive samples show the regions with the normal vessel wall structure have high heat values while the core regions of TCFA have low heat values. It indicates the proposed MVCTN has

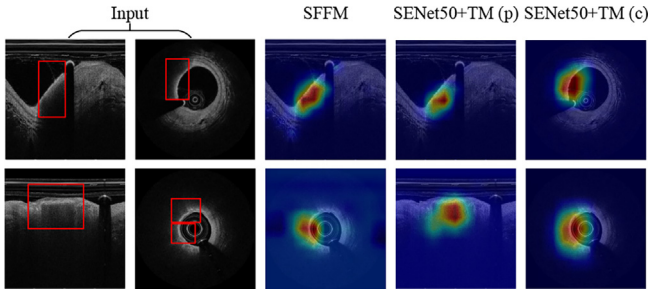

Fig. 9. A CAM-based visualization comparison between SFFM and SENet50 + TM. The first two columns show the polar and Cartesian views of images, respectively. And the rest columns show the CAMs. In addition, the regions of TCFA are indicated with the red box.

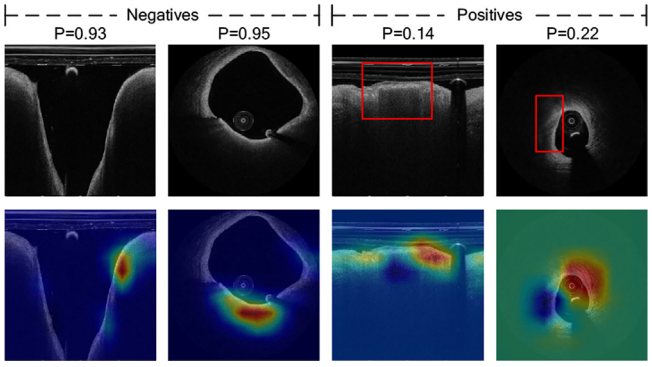

Fig. 10. CAM-based visualization results of the proposed MVCTN for misclassified samples. The first row shows the primary view of images, and the second row shows their CAMs. Besides, the regions of TCFA are indicated with the red box and the predicted positive probabilities are also indicated.

an ability to distinguish the region of TCFA from the region of the normal vessel wall.

5.2. Selection of the Primary View

We further study how to select an appropriate view as the primary view in the proposed MVCTN. First of all, we can find that there is no significant performance difference between MVCTN (p) and MVCTN (c) on the two datasets, especially on OPRD. Secondly, as aforementioned, CAMs of the proposed MVCTN are based on the primary view of images. However, compared to original images, the new images generated by the view transformation have some information added or missing because of the up-sampling or down-sampling operation. It is more straightforward for cardiologists to use the CAMs to analyze the original images. All in all, we argue that it is reasonable to take the original view as the primary view in the proposed MVCTN.

5.3. Clinical Benefits

In clinical routine, hundreds of IVOCT images are acquired from each patient's pullback, which requires automated TCFA identification for fast and accurate decision support. The proposed MVCTN is deployed on one Nvidia GTX Titan X, and processes one image every 0.02 s, which means the proposed MVCTN takes only 4 s to process a patient's pullback consisting of two hundred IVOCT images. This processing speed is sufficient for clinical implementation. Next, it is worth noting that before implementing the view transformation, the CCCV-IVOCT dataset only has the polar view of images and OPRD only has the Cartesian view of images. Therefore, compared to SENet50 trained with the original single view of images, the proposed MVCTN improves AUC, F1-score and recall on the CCCV-IVOCT dataset by 4.4%, 7.9%, and 21.7%, respectively, while improving accuray, F1-score, precision and recall on OPRD by 2.8%, 2.9%, 1.7% and 3.7%, respectively. It is also worth noting that the recall of the proposed MVCTN is better than the precision. The reason might be that the proposed MVCTN is more sensitive to the pathological features in the salient regions due to the presence of the CCSM. However, we do not consider this is a disadvantage. Firstly, the proposed MVCTN aims to fastly and accurately determine the location of TCFA in a patient's pullback. We do not expect to miss any suspicious locations. Secondly, as shown in Fig. 10, the proposed MVCTN misclassifies lipid plaques as TCFA, and the lipid plaques may develop to dangerous plaques, which thus are also worthy of cardiologist's attention. Furthermore, via the CCSM, the proposed MVCTN can mimic the diagnostic process of cardiologists to extract more discriminative features from the salient regions, thereby improving the identification performance. This mechanism may be easily understood and accepted by cardiologists. Lastly, the proposed MVCTN can also use CAM to provide some meaningful visualization. For one thing, it highlights the regions with the normal vessel wall structure as much as possible in the CAMs of the negative samples. For another thing, it highlights the core regions of TCFA in the CAMs of the positive samples. Consequently, the proposed MVCTN will potentially be a useful tool with explainable results for cardiologists, especially for inexperienced cardiologists, to support them in making diagnostic decisions.

5.4. Limitations and Future Work

Although the proposed MVCTN shows excellent TCFA identification performance, it still has several limitations. Firstly, with the memory limitation of the GPUs used in this paper, the proposed MVCTN merely uses one auxiliary view to assist one primary view. However, the primary view and the auxiliary view can simultaneously assist each other. Thus a mutually assisted network based on the proposed MVCTN could be built, and the interaction between the two views of outputs could also be worth exploring. Secondly, the proposed MVCTN is trained with the 2D IVOCT images that are selected from 3D IVOCT volumes. It means that a large number of images are not utilized and using them may further boost performance. Therefore, a 3D extended version considering the relation of adjacent images could be studied. Last but not least, the proposed MVCTN is only used for the TCFA identification in this paper. However, the proposed MVCTN may be applicable to other classification tasks. There are two conditions for such wider applications. The first is that the active contour model can be used to detect salient regions. The second is that two different views of images can be acquired. Therefore, applying the proposed MVCTN to other classification tasks could be investigated.

6. Conclusion

In this paper, we propose a multi-view contour-constrained transformer network (MVCTN) for the TCFA identification in the IVOCT images. The proposed MVCTN can not only extract discriminative features from the salient regions (i.e., the vessel walls) and enhance the visual interpretability based on CAM via the CCSM, but also build the global-range relation between the two views to allow the auxiliary view to assist the primary view at multiple feature scales via the TM. The experimental results on the CCCV-IVOCT dataset and OPRD demonstrate that the proposed MVCTN outperforms the other single-view and multi-view methods. Besides, the proposed MVCTN can also provide some meaningful

CAM-based visualization for cardiologists, especially for inexperienced cardiologists, to support them in making fast and accurate decisions.

CRediT authorship contribution statement

Sijie Liu: Conceptualization, Methodology, Software, Writing - original draft. **Jingmin Xin:** Supervision, Writing - review & editing. **Jiayi Wu:** Conceptualization, Writing - review & editing. **Yangyang Deng:** Investigation. **Ruisheng Su:** Writing - review & editing. **Wiro J. Niessen:** Writing - review & editing. **Nanning Zheng:** Supervision, Funding acquisition. **Theo van Walsum:** Supervision, Writing - review & editing.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700800, and the National Natural Science Foundation of China under Grant 62088102 and 82000336. We thank Professor Bo Yu from Department of Cardiology, Second Affiliated Hospital of Harbin Medical University, Harbin, China for help in building the Optical coherence tomography Plaque Recognition Database (OPRD).

References

[1] H. Sinclair et al., OCT for the identification of vulnerable plaque in acute coronary syndrome, J. Am. Coll. Cardiol. Img. 8 (2) (2018) 198–209.
[2] Frank D. Kolodgie et al., The thin-cap fibroatheroma: a type of vulnerable plaque: The major precursor lesion to acute coronary syndromes, Curr. opin. cardiol. 16 (5) (2001) 285–292.
[3] L.S. Athanasiou, N. Bruining, F. Prati, D. Koutsouris, Optical coherence tomography: basic principles of image acquisition, in: Intravascular Imaging: Current Applications and Research Developments, 2011, pp. 180–193.
[4] F. Prati et al., Expert review document on methodology, terminology, and clinical applications of optical coherence tomography: physical principles, methodology of image acquisition, and clinical application for assessment of coronary arteries and atherosclerosis, Eur. Heart J. 31 (4) (2010) 401–415.
[5] J. Wang, OCT image recognition of cardiovascular vulnerable plaque based on CNN, IEEE Access 8 (2020) 140767–140776.
[6] M. Xu et al., Fibroatheroma identification in intravascular optical coherence tomography images using deep features, in: Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2017, pp. 1501–1504.
[7] N. Gessert et al., Automatic plaque detection in IVOCT pullbacks using convolutional neural networks, IEEE Trans. Med. Imag. 38 (2) (2018) 426–434.
[8] P. Shi, J. Xin, S. Liu, Y. Deng, N. Zheng, Vulnerable plaque recognition based on attention model with deep convolutional neural network, in: Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2018, pp. 834–837.
[9] S. Liu, Y. Deng, J. Xin, W. Zuo, P. Shi, N. Zheng, Srcnn: Cardiovascular vulnerable plaque recognition with salient region proposal networks, in: Proc. 2nd Int. Conf. Graph. Signal Process., 2018, pp. 38–45.
[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2921–2929.
[11] Y. Kim, S. Kim, T. Kim, C. Kim, Cnn-based semantic segmentation using level set loss, in: Proc. IEEE Wint. Conf. Appl. Comput. Vis. (WACV), 2019, pp. 1752–1760.
[12] J.J. Rico-Jimenez et al., Automatic classification of atherosclerotic plaques imaged with intravascular oct, Biomed. Opt. Express 7 (10) (2016) 4069–4085.
[13] C. He, J. Wang, Y. Yin, Z. Li, Automated classification of coronary plaque calcification in OCT pullbacks with 3D deep neural networks, J. Biomed. Opt. 25 (9) (2020) 1–13.
[14] P. Shi, J. Xin, N. Zheng, Weakly supervised vulnerable plaques detection by IVOCT image, in: Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI), 2020, pp. 1983–1986.
[15] Q. Li et al., Research on automatic identification based on IVOCT images of coronary plaque, in: Proc. Opt. Health Care Biomed. Opt. IX, 2019, pp. 9–19.
[16] L.S. Athanasiou et al., A deep learning approach to classify atherosclerosis using intracoronary optical coherence tomography, in: Proc. Med. Imag. 2019: Comput.-Aided Diagn., 2019, pp. 163–170.
[17] X. Ren, H. Wu, Q. Chen, T. Kubo, T. Akasaka, A tissue classification method of IVOCT images using rectangle region cropped along the circumferential direction based on deep learning, in: Proc. Int. Forum Med. Imag. Asia, 2019, pp. 196–202.
[18] M. Xu et al., Automatic image classification in intravascular optical coherence tomography images, in: Proc. IEEE Reg. 10 Conf. (TENCON), 2016, pp. 1544–1547.
[19] P. Zhou, T. Zhu, C. He, Z. Li, Automatic classifcation of atherosclerotic tissue in intravascular optical coherence tomography images, J. Opt. Soc. Am. A 34 (7) (2017) 1152–1159.
[20] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, Int. J. Comput. vis. 1 (4) (1988) 321–331.
[21] T.F. Chan, L.A. Vese, Active contours without edges, IEEE Trans. Image Process. 10 (2) (2001) 266–277.
[22] Y. Niu, L. Qin, X. Wang, Structured graph regularized shape prior and cross-entropy induced active contour model for myocardium segmentation in CTA images, Neurocomputing 357 (2019) 215–230.
[23] C. Ma, G. Luo, K. Wang, Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of MR images, IEEE Trans. Med. Imag. 37 (8) (2018) 1943–1954.
[24] F. Riaz, S. Naeem, R. Nawaz, M. Coimbra, Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images, IEEE J. Biomed. Health Inform. 23 (2) (2019) 489–500.
[25] A. Vaswani et al., Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.
[26] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7794–7803.
[27] A. Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.
[29] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 945–953.
[30] L. Li, T. Jia, T. Meng, Y. Liu, Deep convolutional neural networks for cardiovascular vulnerable plaque detection, MATEC Web Conf. (2019) 02024.
[31] R. Liu et al., Automated detection of vulnerable plaque for intravascular optical coherence tomography images, Cardiovasc. Eng. Tech. 10 (4) (2019) 590–603.
[32] A. Paszke et al., Pytorch: An imperative style, high-performance deep learning library, in: Proc. Adv. Neural inf. Process. Syst. (NIPS), 2019, pp. 8026–8037.
[33] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
[35] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: representation learning, information fusion, and applications, IEEE J. Sel. Top. Signal Process. 14 (3) (2020) 478–493.
[36] Y. Dai, Yifan Gao, TransMed: Transformers advance multi-modal medical image classification, arXiv preprint arXiv:2103.0594, 2021.
[37] T. Chen et al., Multi-view learning with feature level fusion for cervical dysplasia diagnosis, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assis. Intervent. (MICCAI), 2019, pp. 329–338.
[38] X. Chu et al., Twins: Revisiting the design of spatial attention in vision transformers, in: Proc. Adv. Neural inf. Process. Syst. (NIPS), 2021, pp. 9355–9366.

Sijie Liu received the B.E. degree in automation from Chongqing University, Chongqing, China, in 2016. He is currently a PhD student at Xi'an Jiaotong University, Xi'an, China, as well as a visiting PhD student at Erasmus MC, The Netherlands. His current research interests include stroke, fibroatheroma, medical image analysis and computer vision.

Jingmin Xin received the B.E. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1993 and 1996, respectively. From 1988 to 1990, he was with the Tenth Institute of Ministry of Posts and Telecommunications (MPT) of China, Xi'an. He was with the Communications Research Laboratory, Japan, as an Invited Research Fellow of the Telecommunications Advancement Organization of Japan (TAO) from 1996 to 1997 and as a Postdoctoral Fellow of the Japan Science and Technology Corporation (JST) from 1997 to 1999. He was also a Guest (Senior) Researcher with YRP Mobile Telecommunications Key Technology Research Laboratories Company, Limited, Yokosuka, Japan, from 1999 to 2001. From 2002 to 2007, he was with Fujitsu Laboratories Limited, Yokosuka, Japan. Since 2007, he has been a Professor at Xi'an Jiaotong University. His research interests are in the areas of adaptive filtering, statistical and array signal processing, system identification, and pattern recognition.

Jiayi Wu received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2021. During 2017 to 2018, she was a visiting Ph.D student in the Vascular Imaging Lab (VIL) at University of Washington (UW). She is currently an assistant professor at the College of Artificial Intelligence, Xi'an Jiaotong University. Her current research interests include medical imaging analysis and machine learning.

Yangyang Deng received the MBBS degree from Xi'an Jiaotong Univerity, Xi'an, China, in 2013. She is currently working toward the Ph.D degree with the Department of Cardiology, First affiliated hospital of Xi'an Jiatong Univerity, Xi'an, China. Her current research interests include the application of artificial intelligence in cardiovascular disease and medical image Process.

Ruisheng Su is currently a PhD student at Erasmus Medical Center, The Netherlands. He obtained his Master degree in Electrical Engineering from TU Munich. His research interests include stroke, medical image and video analysis, image guided intervention and computer vision.

Wiro Niessen is full professor in Biomedical Image Analysis and Machine Learning at Erasmus MC and Delft University of Technology, The Netherlands. His interest is in the development and validation of quantitative biomedical image analysis methods, and linking imaging and genetic data for improved disease diagnosis and prognosis, using machine learning (AI). Wiro Niessen is fellow and was president of the MICCAI Society, and is CTO of Health-RI, which aims to develop a national health data infrastructure for reuse of data for research and innovation. In 2015 he received the Simon Stevin award, the largest prize in the Netherlands in Applied Sciences. In 2005 he was elected to the Dutch Young Academy and in 2017 he was elected to the Royal Netherlands Academy of Arts and Sciences. In 2012 he founded Quantib, an AI company in medical imaging where he currently acts as scientific lead.

Nanning Zheng graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975, and received the M.S. degree in information and control engineering from Xi'an Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He joined Xi'an Jiaotong University in 1975, and is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems. Dr. Zheng became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition. He also serves as an executive deputy editor of the Chinese Science Bulletin.

Theo van Walsum is associate professor at the Erasmus MC (Biomedical Imaging Group Rotterdam), and is heading the "Image Guidance in Interventions" research line. His work focuses on improving image guidance by integrating pre-operative image information in various interventional procedures. Challenges addressed are the modeling and tracking of motion and deformation of the anatomy, and the instruments. This research also includes the use of augmented reality devices to integrate information directly in the field of view of the clinician. Additionally, he works in the cardiovascular image processing, specifically on quantitative imaging biomarkers for the heart and coronaries (mainly CT-based) and for stroke (NCCT, CTA, DSA). His interests are medical imaging, image guidance and navigation, visualization and software engineering.