Neural Networks 128 (2020) 172–187

Contents lists available at ScienceDirect

Neural Networks

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

2020 Special Issue

Sequential vessel segmentation via deep channel attention network

Dongdong Hao [a,1], Song Ding [b,1], Linwei Qiu [c], Yisong Lv [d], Baowei Fei [e], Yueqi Zhu [f], Binjie Qin [a,*]

[a] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Cardiology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China
[c] School of Astronautics, Beihang University, Beijing 100191, China
[d] School of Continuing Education, Shanghai Jiao Tong University, Shanghai 200240, China
[e] Department of Bioengineering, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA
[f] Department of Radiology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai Jiao Tong University, 600 Yi Shan Road, Shanghai 200233, China

a r t i c l e i n f o

Article history:
Available online 13 May 2020

Keywords:
X-ray coronary angiography
Deep learning
Vessel segmentation
temporal–spatial features
Channel attention blocks
Class imbalance

a b s t r a c t

Accurately segmenting contrast-filled vessels from X-ray coronary angiography (XCA) image sequence is an essential step for the diagnosis and therapy of coronary artery disease. However, developing automatic vessel segmentation is particularly challenging due to the overlapping structures, low contrast and the presence of complex and dynamic background artifacts in XCA images. This paper develops a novel encoder–decoder deep network architecture which exploits the several contextual frames of 2D+t sequential images in a sliding window centered at current frame to segment 2D vessel masks from the current frame. The architecture is equipped with temporal–spatial feature extraction in encoder stage, feature fusion in skip connection layers and channel attention mechanism in decoder stage. In the encoder stage, a series of 3D convolutional layers are employed to hierarchically extract temporal–spatial features. Skip connection layers subsequently fuse the temporal–spatial feature maps and deliver them to the corresponding decoder stages. To efficiently discriminate vessel features from the complex and noisy backgrounds in the XCA images, the decoder stage effectively utilizes channel attention blocks to refine the intermediate feature maps from skip connection layers for subsequently decoding the refined features in 2D ways to produce the segmented vessel masks. Furthermore, Dice loss function is implemented to train the proposed deep network in order to tackle the class imbalance problem in the XCA data due to the wide distribution of complex background artifacts. Extensive experiments by comparing our method with other state-of-the-art algorithms demonstrate the proposed method's superior performance over other methods in terms of the quantitative metrics and visual validation. To facilitate the reproductive research in XCA community, we publicly release our dataset and source codes at https://github.com/Binjie-Qin/SVS-net.

1. Introduction

1.1. Motivation

Nowadays, cardiovascular diseases have seriously threatened more and more people's health (Townsend et al., 2016). Percutaneous coronary intervention, as the minimally invasive method for cardiovascular disease treatment, has been widely adopted in the clinic. During this intervention, contrast agents are injected into the vessels through one catheter and then X-ray coronary angiography (XCA.[2]) is employed to help surgeons navigate the catheters (Albarqouni, Fotouhi, & Navab, 2017; Jin, Li, Jiang, & Qin, 2017; Yang, Wang, Liu, Tang, & Chen, 2009). With the help of the contrast-enhanced images, doctors can diagnose the coronary artery disease and evaluate therapeutic effects. It is important to accurately and quickly segment vessels from XCA data for the diagnosis and intervention of cardiovascular diseases.

Although vessel segmentation has always been a hot spot due to its significance and complexity in clinical practice (Jin, Hao, Ding, & Qin, 2018; Kerkeni, Benabdallah, Manzanera, & Bedoui, 2016; Soares, Leandro, Cesar, Jelinek, & Cree, 2006), coronary artery vessel segmentation still remains highly challenging because of the poor visual quality of XCA, which is caused by the low

* Corresponding author.
E-mail address: bjqin@sjtu.edu.cn (B. Qin).

[1] The co-first authors contributed equally to this work.

[2] The notations used in this paper are listed in Table 1.

https://doi.org/10.1016/j.neunet.2020.05.005
0893-6080/

Table 1
The main notations used in this paper.

Notation	Explanation
XCA	X-ray coronary angiography
SVS-net	sequential vessel segmentation deep network
CRF	conditional random field
FFO	feature fusion operation
CAB	channel attention block
Conv3D	3D convolution
Block3D	3D residual convolutional block
Conv2D	2D convolution
Block2D	2D residual convolutional block
BN	batch normalization
ReLu	rectified linear unit
sigmoid	sigmoid activation function
ks	convolutional kernel size
LF	low-stage feature maps
HF	high-stage feature maps
DR	detection rate
P	precision
CE	cross entropy
GVEs	gland volume errors

contrast and high Poisson noise of low dose X-ray imaging, overlap of background structures and foreground vessels, complex motion patterns and disturbance of spatially distributed noisy artifacts. Currently most 3D/2D vessel segmentation methods are proposed to segment the vessels from computed tomography angiography, magnetic resonance angiography and a single 2D image (Jin et al., 2018; Moccia, Momi, Hadji, & Mattos, 2018), of which there is no serious disturbance from the noise and overlapping background structures. To obtain the vascular structures from XCA, a few computer vision and machine learning related methods have been developed. Kerkeni et al. (2016) propose an iterative region growing algorithm to integrate both vesselness and direction information in the multi-scale space. However, it fails to recognize thin and peripheral vessel in the low contrast XCA images. Jin et al. (2018) extract the contrast-filled vessels via robust principal component analysis and combine both local and global thresholds to refine vessel segmentation mask (Unberath, Aichert, Achenbach, & Maier, 2017) but with some residuals remained around vessel regions. Felfelian et al. (2016) detect coronary artery regions of interest based on Hessian filter and identify vessel pixels by flux flow measurements. Nevertheless, a series of postprocessing should be performed to improve the robustness and accuracy of segmentation mask.

With the development of neural network-based deep learning, Nasr-Esfahani et al. (2016) use convolutional neural network (CNN) with fully-connected layers to perform vessel segmentation, which overlooks the structure information and temporal correlation in XCA sequence images. To alleviate these issues, Fan et al. (2018) develop a multichannel fully convolutional neural network with live image and corresponding dense matching mask image inputted to the network. However, it should collect corresponding mask images and perform dense matching in advance to segment vessel structure from live images, which is not practical in clinical applications. Most of XCA segmentation algorithms are dependent on the pixels of local windows in a single frame of XCA sequences, so that they waste lots of temporal–spatial contextual information in XCA sequences, which can be important to infer whether the pixels belong to the foreground vessel regions or not. Although current vessel segmentation (Moccia et al., 2018) methods have made great progress in segmentation accuracy, they are still inefficient in the large dynamic datasets from the complex XCA sequences with many noisy and overlapped background artifacts.

To design a robust and efficient XCA segmentation algorithm for clinical applications, we should have a good knowledge about the XCA images' characteristics. Usually, with the illumination of X-rays at specific angle or direction, various 3D anatomical structures such as vessels, lungs, spines, diaphragms and bones are projected along definite path and displayed as overlapped 2D structures on the X-ray angiogram plane. To simplify description, we straightly identify vessels as foreground and regard other overlapped structures as background. Low dose radiopaque contrast agents are primarily injected into angiocarpy to enhance the visibility of vessels in XCA images. Even so, the vessels in XCA images are still of poor visibility due to the following factors: (1) The projection onto 2D plane causes overlap of adjacent tissues. Therefore, foreground vessel regions are badly disturbed by respiratory motion (Blondel, Malandain, Vaillant, & Ayache, 2006). Moreover, it is very difficult to differentiate foreground vessels from background due to the low intensity contrast between the vessels and the background in low-dose X-ray imaging (Xia et al., 2019); (2) Vessels usually have plenty of branches. Radiopaque contrast agents flow at different speeds in each branches. As a result, different vessels branches vary in gray values and some vessel regions cannot be clearly visible in the same time (Qin et al., 2019); (3) The spatially distributed Poisson noises (Zhu et al., 2013) caused by low-dose X-ray imaging reduce the SNR between the signals and noise. The noisy background artifacts and foreground vessels have different motion patterns, so that these noisy and dynamic structures severely interfere with the feature extraction and classification for vessel segmentation. All abovementioned difficulties determine that sequential vessel segmentation from XCA image sequences is a highly challenging task.

Hierarchical deep CNN features have proven incredibly effective at a wide range of image classification and image segmentation tasks. The generic deep CNN feature extractor trained for general purpose image segmentation often perform surprisingly well for novel segmentation tasks without doing any fine-tuning beyond training a linear classifier (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2018; Ronneberger, Fischer, & Brox, 2015). This success is often explained by the built-in invariance of deep CNN features to local image transformation and the insensitivity of deep CNN features to shading, low-contrast, etc. We might hope that these invariances would prove useful in our challenging setting of sequential vessel segmentation. However, our problem differs in that we need to segment sequential foreground vessels from the noisy and overlapped background with similar appearances rather than simply training a k-way classifier. To overcome all the mentioned issues of XCA segmentation by deep network, we give the following specific considerations:

(1) although unsupervised learning or weakly-supervised learning (Huang, Change Loy, & Tang, 2016; Kallenberg et al., 2016) with deep CNN features have developed a lot, they still fail to obtain competitive performance compared with supervised ones. That is because supervised learning introduces straight priors to guide the learning process. In view of the requirement of segmentation accuracy, we adopt supervised learning strategy. As a data-driven method, supervised deep learning depends on large annotated training datasets to ensure excellent performance especially for the video related tasks. Unfortunately, there is no readily available public dataset for vessel segmentation from XCA sequence. To this end, we have collected many XCA sequences from our university-affiliated hospitals and employed several clinical experts to annotate vessel label so that we can set up ground truth for vessel segmentation.

(2) Recent approaches have concentrated on some but not all the fore-mentioned issues and try to make use of temporal information for vessel segmentation. They either take adjacent multiple frames in a sliding window centered at current frame as whole input straightly (Hao, Ma, & van Walsum, 2018) or make

pre-matching to generate segmentation mask (Fan et al., 2018; Khanmohammadi, Engan, Eftest, Sœland, Larsen, et al., 2017). The former indeed introduces not only temporal information but also much disturbances; the latter needs extra dense image matching which is time consuming and incorrect especially for the low contrast images. Recent video segmentation methods explore how to properly utilize the temporal information in the sequential images, *i.e.*, estimate optical flow (Rashed, Yogamani, El-Sallab, Krizek, & El-Helw, 2019; Sun, Yang, Liu, & Kautz, 2018) for modeling the motion among adjacent frames, apply convolutional LSTMs (Pfeuffer & Dietmayer, 2019; Pfeuffer, Schulz, & Dietmayer, 2019) to learn long short-term dependencies in video sequence. They learn effective temporal–spatial consistent features in natural scene image, however they may exist as data matching errors (Simoncelli, Adelson, & Heeger, 1991) when applied in the noisy, low-contrast, and blurry XCA medical images. Therefore, how to design an effective network that can learn proper temporal–spatial vessel features from the noisy background will be most important for our work.

(3) The class imbalance problem caused by the imbalance ratio between the number of foreground vessel pixels and background pixels typically lies in the challenging vessel segmentation tasks and must be well treated to boost the vessel segmentation. Current methods partly addressed this issue by weighted cross entropy (Lim & Keles, 2018) or proper training patch selection strategy (Nasr-Esfahani et al., 2016; Yan, Yang, & Cheng, 2018). However, they failed to completely solve that imbalance problem. Inspired by the work in Ambrosini, Ruijters, Niessen, Moelker, and van Walsum (2017), Zhang et al. (2018), we utilize Dice loss function to guide the network learn balanced information representation between foreground vessel and background pixels. In addition, the deeper the network, the stronger is the representation capacity. However, the optimization of deep network structure is extremely difficult due to the common problem of gradient vanishing and gradient explosion in deep network. We have properly integrated residual blocks (He, Zhang, Ren, & Sun, 2016; Szegedy, Ioffe, Vanhoucke, & Alemi, 2017; Xie, Girshick, Dollár, Tu, & He, 2017) into our vessel segmentation deep network to alleviate the above problem.

In summary, this work has the following contributions:

(1) We propose an encoder–decoder-based sequential vessel segmentation deep network architecture called **SVS-net** that acquires the temporal–spatial information from the several contextual frames in a sliding window centered at current frame to segment the 2D vessel masks of the current frame in XCA sequence: (i) In encoder network, temporal–spatial vessel features from the complex and noisy background artifacts are extracted in 3D (2D+t) manners; (ii) The extracted features are then fused along temporal axis in the skip connection layers, which transform the contextual 3D temporal–spatial feature maps into 2D spatial feature maps for the segmentation of current frame. This 3D-2D fusion introduces dimension reduction to further help reduce the subsequent calculation burden and trainable parameters; (iii) Finally, the decoder network efficiently integrates the fused temporal–spatial information in XCA image sequence by feature refinement and subsequently decodes the refined features in 2D ways to produce the segmented vessel masks. Specifically, a channel attention mechanism is implemented in channel attention blocks (CABs) to refine the fused temporal–spatial features by adaptively highlighting and learning the discriminative vessel features from the noisy background artifacts via weighting the feature maps. To the best of our knowledge, it is the first time to apply channel attention mechanism in taking both temporal and spatial information into the deep sequential vessel segmentation architecture. Moreover, the proposed SVS-net can be trained in an end-to-end way.

(2) We publicly release a XCA database with ground truth annotation. The lack of XCA data with annotated label impedes the further exploration on XCA related researches such as vessel segmentation and vessel recovery in deep learning community. Therefore, we established database to promote these studies with detailed data description in the method section of this paper.

(3) We employ Dice loss function in deep network to alleviate the severe class unbalance problem in sequential vessel segmentation and validate its significance when compared with binary cross entropy. We have evaluated the effectiveness of 3D temporal–spatial features and CABs used in the proposed SVS-net by comparing them with 2D counterparts and other state-of-the-art methods. Extensive experiments have verified SVS-net's superior performance over other algorithms.

1.2. Related works

This section introduces the recent works related to vessel segmentation. Vessel segmentation algorithms can be simply divided into two categories: traditional segmentation methods and deep learning-based methods. Recent traditional methods and deep learning-based methods are summarized in this section respectively.

1.2.1. Traditional vessel segmentation methods

Various traditional approaches have emerged in the past decades, including filtering based methods (Chaudhuri, Chatterjee, Katz, Nelson, & Goldbaum, 1989; Frangi, Niessen, Vincken, & Viergever, 1998; Moccia et al., 2018; Soares et al., 2006), tracking based algorithms (Kumar, Vázquez-Reina, & Pfister, 2010; Staal, Abràmoff, Niemeijer, Viergever, & Van Ginneken, 2004), and model-based methods (Chen, Zhang, & Cohen, 2019; Dehkordi, Hoseini, Sadri, & Soltanianzadeh, 2014; Law & Chung, 2009). Filtering-based methods develop specific filters convolving with the original images to enhance the tubular structures (Chaudhuri et al., 1989; Frangi et al., 1998; Moccia et al., 2018). In Chaudhuri et al. (1989), the intensity profile of the vessel was approximately modeled as a Gaussian shaped curve and then 12 different matched filtering templates are utilized to search for the latent vessel segments along different directions. Frangi et al. (1998) propose a common vesselness enhancement technique, where the second order derivative is calculated to form Hessian matrices and the corresponding eigenvalues are analyzed.

Different from above filtering based approaches, other classes of filters are developed to extract vessel features, such as the ridges feature, the Radon-like features and Gabor wavelet features (Kumar et al., 2010; Soares et al., 2006; Staal et al., 2004) and construct pixel-wise vessel feature descriptors for classification. Although these methods enhance vessel structure to some degree, they are executed with high time complexity for the pixel-wise manipulation. Besides, they usually serve as the preprocessing step and further postprocessing like threshold methods and morphology operation should be utilized to construct final refined vessel masks.

In regard to tracking based segmentation methods, the initial seed points on the vessel edges are chosen firstly and then the tracking process starts under the guidance of image-derived constraints. The tracking algorithms vary from each other according to the different definition of the tracking constraints. For example, Makowski et al. (2002) employ two-phase based method during vessel extraction, which use balloon segmentation and snake

segmentation, respectively. Recursive tracking (Carrillo, Hoyos, Dávila, & Orkisz, 2007) is performed by accumulating pixels on the basis of a cluster algorithm with geometry and intensity constraints, while level set evolution (Manniesing, Viergever, & Niessen, 2007) is employed to track the vessel axis with the evolution process being guided by imposing shape constraints on the skeleton topology. However, these tracking based methods fail to segment out small vessels from the complex and overlapped noisy background, and human intervention is needed to set and adjust the algorithms' parameters.

Usually, model-based methods are designed on the basis of specific shapes and appearance of the interested structures (Chen et al., 2019; Moccia et al., 2018) contained in the images. There are mainly three categories including parametric model, deformable model-based segmentation, and statistic model. We can refer to Moccia et al. (2018) for detailed introduction. The model-based methods still have many unsolved problems on detecting small vessels, finding out right parameters to fit the model, and recognizing abnormalities consisted in the diseased vessels. Overall, above traditional segmentation methods require professional knowledge to elaborately construct feature engineering and the complex processing procedures and their segmentation accuracy and real time performance still need to be improved.

1.2.2. Deep learning-based methods

Compared with traditional segmentation methods, deep learning ones automatically learn proper feature representation and perform better on generalization capacity as well as inference speed. Consequently, deep learning methods can earn a top rank in many computer vision fields including segmentation, detection, classification and so on (Sakkos, Ho, & Shum, 2019; Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018).

Recently, CNN-based methods have been broadly applied to medical image segmentation such as retinal vessel segmentation (De Fauw et al., 2018; Yan et al., 2018). Generally, the works in Liskowski and Krawiec (2016), Nasr-Esfahani et al. (2016) treat the retinal vessel segmentation task as binary classification, in which a typical classification network containing several stacked convolutional layers and three fully-connected layers are adopted. To alleviate the limitation of annotated data, a patch-based learning strategy is implemented. However, there exist several problems: (1) The limited size of patch means a limited receptive field, which fails to provide sufficient contextual information for accurate segmentation. Fusing predictions of all patches in the image to form the final vessel mask needs to run the whole network many times and is very time consuming. (2) Fully-connected layers function as feature weighting and fuse both local and global information from feature space to label space. However, they contain almost 80% the parameters of the whole network, which may result in overfitting (Ruder, Dosovitskiy, & Brox, 2018; Wen, Zhang, Li, & Qiao, 2019). (3) Due to the localization requirement, fully-connected layers overlook the spatially structured features that are significant for segmentation tasks. Furthermore, the usage of fully-connected layers sets a limit on the network's input size.

To deal with these inherent problems, fully convolutional network (FCN) (Dasgupta & Singh, 2017; Maninis, Pont-Tuset, Arbeláez, & Van Gool, 2016) is proposed to replace fully-connected network in segmentation tasks. Recently, a FCN based on encoder–decoder architecture is introduced in Fan et al. (2018), which adopts a two-channel input strategy and largely depends on the pre-matching between the two-channel inputs. Mo and Zhang (2017) combine some intermediate layers' outputs and fuse hierarchical features together to set up the final segmentation map. Similarly, a deeply supervised multi-level and multi-scale network with short connections is utilized to ease the gradient back propagation for retinal vessel segmentation (Guo, Gao, Wang, & Li, 2018). However, proper feature fusing weights should be carefully set. Fu, Xu, Wong, and Liu (2016) have modeled the retinal vessel segmentation as a pixel-level classification based on modified FCN. Unfortunately, the lack of smoothness constraint and the limited receptive fields in FCN result in false positive (spurious) regions in segmentation output. Therefore, conditional random field (CRF) formulating long-range interactions between pixels is employed to refine the coarse vessel maps (Hu et al., 2018). However, most vessel segmentation algorithms are proposed for solely segmenting vessels from 3D and/or 2D vessel images, which are not appropriate to confront the poor visual quality as well as complex and dynamic background artifacts in sequential vessel segmentation of XCA sequences.

To focus on the most salient features and suppress the less relevant artifacts simultaneously during learning, attention mechanism equipped within deep learning network (Chen et al., 2017; Hu, Shen & Sun, 2018; Jetley, Lord, Lee, & Torr, 2018) is widely adopted for various tasks including image classification (Peng, He, & Zhao, 2018; Schlemper et al., 2019; Wang et al., 2017), image segmentation (Kearney et al., 2019; Li, Dong, Du, & Mu, 2019; Schlemper et al., 2019; Yu et al., 2018) and object detection (Fu, Zhao, & Gu, 2018; Li & Yu, 2018; Li et al., 2016). Attention mechanism is derived from the study of human visual mechanisms, with which people usually pay more attention to the most salient information while neglect some trivials. The key idea of attention mechanism lies in properly generating attention maps to weight feature maps which are extracted by convolutional layers. Zhou, Khosla, Lapedriza, Oliva, and Torralba (2016) use fully convolutional networks and utilize global average pooling to generate attention maps. Hu, Shen et al. (2018) and Yu et al. (2018) have proposed channel attention mechanism to obtain weight vectors by modeling the channel-wise relationship between different feature maps. Chen et al. (2017) integrate both spatial and channel-wise attention in CNN for image captioning. However, to the best of our knowledge, there are no deep network utilizing the channel attention mechanism to extract most salient vessel features from complex and dynamic background artifacts in spatial–temporal contexts for XCA vessel segmentation.

2. Methods

2.1. Overview

The architecture is equipped with temporal–spatial feature extraction in encoder stage, feature fusion operation (FFO) in skip connection layers and CAB in decoder stage. In the encoder stage, a series of 3D convolutional layers are employed to hierarchically extract temporal–spatial features. Skip connection layers subsequently fuse the temporal–spatial feature maps and deliver them to the corresponding decoder stages. To learn discriminative feature representation and suppress the complex and noisy artifacts in the XCA images, the decoder stage effectively utilizes CAB to refine the intermediate feature maps from skip connection layers.

In the proposed SVS-net, (1) we introduce 3D residual blocks (see Fig. 1) to extract multi-scale temporal–spatial features while ease network optimization in feature encoder stage; (2) these 3D features are integrated by the skip connection layers (see Fig. 1), which fuse the temporal–spatial 3D feature maps along temporal axis and generate the fused 2D spatial feature maps. Through the fusion at the left bottom of Fig. 1, the feature maps' dimension mismatch problems between the 3D encoder stage and the 2D decoder stage are solved and the computation cost is also reduced; (3) the fused features are passed to CAB (see the right bottom of Fig. 1 to refine the vessel features from the noisy background and then transmitted to the decoder stage.

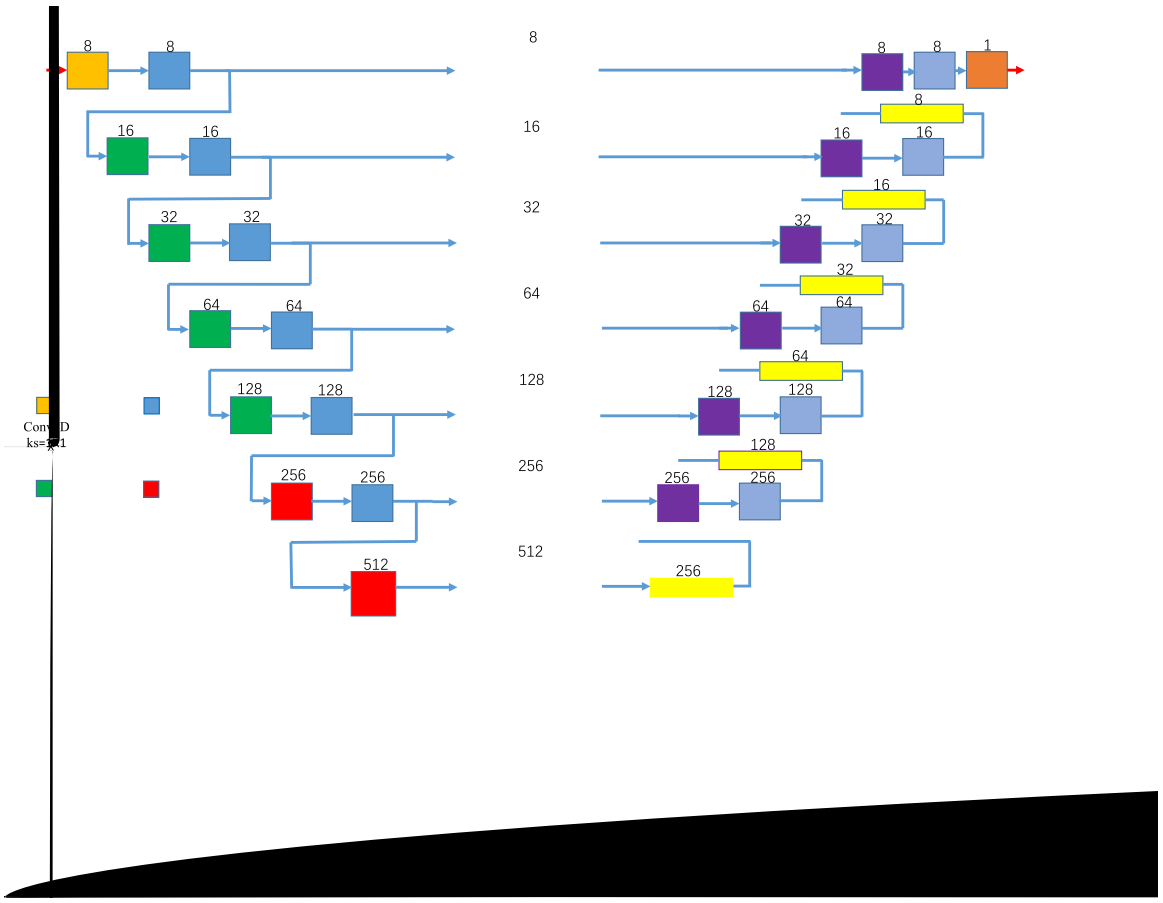

Fig. 1. The proposed network architecture is based on U-net with the encoder network extracting 3D feature from the input sequence and the decoder network learning the salient feature via upsampling and operation of CAB, between the encoder and decoder network is the skip connection layers with FFO. The numbers 8, 16, 32, …, above each block denoting the number of output channels for that block. Convolutional kernel sizes and strides (s: strides) for each block are given in legend. In the FFO and CAB at the bottom, the $F \in \mathbb{R}^{C\times T\times H\times W}$ denotes the temporal–spatial feature maps, C denotes channel axis, T denotes temporal axis, H denotes height axis, W denotes width axis, $F^c \in \mathbb{R}^{T\times H\times W}$: the cth channel of temporal–spatial feature maps. $F^c_{fusing} \in \mathbb{R}^{H\times W}$ denotes the cth channel of fused temporal–spatial feature map through Conv3D with kernel size 4 × 1 × 1 and strides (1,1,1).

(4) Furthermore, Dice loss function is implemented to train the proposed deep network in order to tackle the class imbalance problem in the XCA data due to the imbalanced ratio between background pixels and foreground pixels. CAB in the decoder stage and FFO in the skip connection layers used in the proposed architecture are also displayed at the bottom of Fig. 1.

In the following part of this section, we illustrate the architecture and its training setup in detail. Data augmentation methods and loss function are also introduced in this section.

2.2. Experimental setup

The XCA image sequence consists of a set of frames (F_1, F_2, …F_n). Each frame F_i corresponds to a binary probability map Y_i where the value of the foreground vessel pixels is 1 and the other background regions is 0. For intuitive perspective, single frame fails to provide enough contextual information to infer one pixel belonging to foreground or background because of the low contrast of intensity and the similar appearance between the foreground and background. Among successive frames, contrast-filled vessel regions move fast and consistently through the contiguous frames and the noisy and dynamic background artifacts fluctuate synchronously along with human breathing and heart beating. Therefore, these consistent contexts can serve as the auxiliary temporal–spatial information to accurately identify vessels from background. In this work, we use successive 4 frames (*i.e.*, $F_{i-2}, F_{i-1}, F_i, F_{iC1}$) as input to generate predicted probability map (*i.e.*, P_i) with considering that too many frames will increase the burden of memory and calculation. Furthermore, due to the salient motion disturbances introduced by heart beating and breathing in a relative long period, too many frames will result in big differences of the vessel's shapes and positions between the first and last frames, causing the temporal–spatial contexts turning into misleading information.

To verify the rationality of the input configuration of 4 frames, we respectively input successive frames, *i.e.*, 2, 3, 4, 5 frames, into the network to investigate the network convergence performance. As shown in Fig. 2, there are slight differences in terms of the convergence results of loss function (DiceCoef) in the training set. The smaller that the loss becomes, the better the fitting performance that the model achieves. When we input 4 frames into the network, the loss converges at about −0.86, which is the smallest compared with other input strategies. Therefore, 4 frames are reasonable and feasible to the input configuration for accurate segmentation results. We further explore hyper-parameter setup in a sensitivity analysis of the parameters including the learning rate and the size of input images. In our baseline model, we simply set up learning rate and input size as 0.01 and 512 × 512 respectively. In the subsequent contrast experiments, we merely change either learning rate (*i.e.*, 0.1, 0.001) or input size (*i.e.*, 128 × 128, 256 × 256) and make other experimental configurations remain unchanged. As shown in Fig. 3, the learning rate and input size make a big difference to the training loss. The smaller input size means the limited receptive field so that

Fig. 2. The training loss curve (left) and its local enlarged curve (right) for different input strategies in training process. The 4 frame's input strategy can achieve the least training loss.

Fig. 3. The hyper-parameter experiments (left) and its local magnified curve (right) comparing different learning rates (lr) and input sizes with baseline (lr: 0.01, input size: 512 × 512.).

the model suffers from performance degradation. Furthermore, the smaller initial learning rate may cause the model to get stuck into local minima in optimization, which also decreases the segmentation performance. By these experiments, we can conclude that our settings of learning rate (0.01) and the input size (512 × 512) lead to the best performance considering both speed and accuracy.

2.3. Modified U-net architecture

We employ U-net as the fundamental architecture in SVS-net. U-net (Ronneberger et al., 2015) is a classical and powerful segmentation network architecture widely used for biomedical images by effectively exploring the underlying high-resolution and low-resolution information in biomedical image. Using skip-layers to build a bridge transmitting multi-scale information from encoder network to decoder network, U-net can improve the spatial accuracy of a deep CNN for final high-resolution segmentation results.

On the one hand, XCA images have low contrast and fuzzy boundary, which require more high-resolution detail information for accurate segmentation. The skip connection mechanism in U-net allows high-resolution information delivery to the decoder network for detail recovery. On the other hand, the internal tissue structures with their topologies in XCA images are relatively fixed, the distribution of segmentation targets in the XCA images is regularly presented with simple and clear semantics, which require more low-resolution information to present accurate semantic information for the target object recognition.

Multiple downsampling operations in U-net's encoder network appropriately provide low-resolution information for contextually semantic recognition. In Fig. 1, the encoder network captures 3D temporal–spatial contexts through 3D convolutions followed by 3D residual convolutional blocks except the last convolutional operation. The decoder network enables precise localization of high-resolution target vessel semantic information via upsampling layers and CAB.

Aiming at the accurate 2D+t XCA vessel segmentation, we make following adaptations based on conventional U-net: (1) In the encoder network, there are 7 stages of 3D convolution. The first six convolutional stages followed by 3D residual convolutional block (in Section 2.4) are utilized to extract rich temporal–spatial feature representations, which provide contexts for subsequent vessel mask inference in the decoder network. The output of each 3D residual convolutional block is passed to the next 3D convolutional stage and the skip connection layer respectively. At the last two 3D convolutional stages, spatial dropout (0.5) is employed before executing convolution to avoid overfitting; (2) In the skip connection layers, we fuse temporal–spatial feature representation by mapping from 3D space to 2D space via 4 × 1 × 1 convolutional kernel in FFO, where the first dimension of convolutional kernel indicates the temporal axis, *i.e.*, 4 indicating 4 channels (frames) in the temporal domain. The temporal domain features are then fused together by temporal axis convolution. The FFO can be formulated as follows:

$$X_{F_l} \; \mathrm{D} \; Squeeze(X_F \quad W) \tag{1}$$

where $X_F \;\; 2 \;\; \mathbb{R}^{C\times T\times H\times W}$ is the spatial–temporal feature map coming from the output of each 3D convolutional stage in the encoder network, $X_{F_L} \; 2 \; \mathbb{R}^{C\times H\times W}$ denotes fusing spatial–temporal feature map, C, T, H, W are the features' channel dimension, temporal dimension, height, and width, respectively. W denotes 4 × 1 × 1 convolutional kernel, represents convolution operation, Squeeze denotes dimension compress, a straightforward schematic can be seen at the left bottom of Fig. 1; (3) In the decoder network, to gradually recover the feature maps' spatial resolution, we take advantage of the parameter-free bilinear upsampling strategy rather than transposed convolutional operations, which contributes to reduce the number of trainable parameters without degrading the segmentation performance (De Fauw et al., 2018). Each upsampling layer is followed by one CAB (see the right bottom of Fig. 1) and one 2D residual convolutional block (Block2D, see Fig. 1). Note that the high-stage and low-stage feature map outputs with the same resolution from the upsampling layer and the skip connection layer are inputted simultaneously to CAB (as illustrated at the right bottom of Fig. 1), which is employed to learn the most discriminative features from noisy and complex background artifacts (see the details in Section 2.5). After the last 2D residual convolutional block, we employ 1 × 1 convolution followed by sigmoid activation function to yield the final vessel mask.

2.4. 2D and 3D residual convolutional blocks

Generally speaking, increasing the depth of networks can improve network generalization capacity. However, a very deep network implies the difficulty in promoting gradient back propagation, which results in the poor performance. To overcome this problem, He et al. (2016) develop the deep residual network to facilitate gradient back propagation by identity mapping connection. Zagoruyko and Komodakis (2016) demonstrate that the two stacked convolutional layers in single residual block is optimal architecture compared with other settings. Hence, we follow the strategy as advised in Zagoruyko and Komodakis (2016) and employ 3D residual blocks and 2D residual blocks in encoder and decoder networks respectively.

2.5. Channel attention mechanism

To learn more rich and multi-scale feature representation for extracting vessels from complex and dynamic background artifacts, the proposed SVS-net firstly extracts multiple types of features by multiple convolutional kernels in every convolutional stage of the encoder stage (see Fig. 1). Note that there exist three problems: (1) each channel of feature maps represents one specific feature type but not all features are equally significant to the final output; (2) XCA sequence contains not only the target vessels but also much disturbance of overlapping structures that have similar appearances and intensities to vessels, these disturbances are aggregated nearly at different positions with their relatively various moving speeds. Therefore, these disturbances are distributed in different feature channels; (3) Note that the skip connection layers fuse the temporal–spatial features through 4 × 1 × 1 convolution, which performs linear combination in temporal domain. This combination inevitably introduces extra noisy artifacts into different feature channels besides the noisy disturbances inherent in the XCA sequences. As shown in Fig. 4(a2), (b2), the fused spatial feature map contains a lot of noise artifacts from the background area, which may decrease the accuracy of vessel detection. Therefore, the fused spatial feature map from the output of skip connection layer should be well treated to weaken the noise disturbance from the noisy backgrounds and emphasize the vessel feature simultaneously. To this end, we introduce an effective scheme called as channel attention mechanism for highlighting foreground vessel features and noise removal.

Through the operation of CAB, the SVS-net can adaptively highlight some channel information meanwhile suppress the trivial channel information. Hence, the predicted probability map is gradually improved. Inspired by the works (Hu, Shen et al., 2018; Yu et al., 2018), we introduce the CAB to weight the feature maps from the low-stage output from the skip connection layer and then combine with the corresponding high-stage feature maps that are outputted from the upsampling layer. High-stage output feature maps contain more advanced global semantic information while low-stage feature maps contain more detailed yet noisy information, therefore the high-stage features can provide clues to screen useful information from low-stage feature maps and generate more pure feature representation. Under the guidance of high-stage features, the attention weights are learned and used to obtain discriminative salient features. As shown in Fig. 4(a3), (b3), the low-stage feature map from the output of skip connection layer is refined by the CAB. From Fig. 4(a2)–(a3) and Fig. 4(b2)–(b3), the background noises in Fig. 4(a2), (b2) are greatly reduced while the foreground vessel features are highlighted in Fig. 4(a3), (b3).

Specifically, the CAB do the following operations (see the right bottom of Fig. 1): the low-stage feature maps $X_{F_l} \;\; 2 \;\; \mathbb{R}^{C\times H\times W}$ and the corresponding high-stage feature maps $X_{F_h} \;\; 2 \;\; \mathbb{R}^{C\times H\times W}$ are concatenated together to make feature maps $X_F \; 2 \; \mathbb{R}^{2C\times H\times W}$. Furthermore, a global average pooling is performed on the concatenated feature maps to generate the weighted vector $W_{X_F} \;\; 2 \;\; \mathbb{R}^{2C\times 1\times 1}$. (Yu et al., 2018). Two 1 × 1 convolutional operations, which are followed by the rectified linear unit function and sigmoid function, respectively, are performed on $W_{X_F} \;\; 2 \;\; \mathbb{R}^{2C\times 1\times 1}$ to learn inter-channel relationship and the final channel attention weights vector $W_{X_{F_l}} \;\; 2 \;\; \mathbb{R}^{C\times 1\times 1}$ is achieved. The obtained attention vector multiplies low-stage feature maps in channel-wise manner, then the weighted feature maps from low stage are added with the corresponding high-stage feature maps to be subsequently passed to the next layer. The whole process of generating attention weights can be expressed as:

$$W_{X_{F_l}} \; \mathrm{D} \; \phi(\varphi(GAP(X_F))) \tag{2}$$

Fig. 4. Two instances of feature visualizations for illustrating the CAB's effects: suppresses the noises in the background areas while highlights the foreground vessel feature. From left to right, each row displays the original XCA image; the 2nd channel of fused spatial feature maps in the output of the second skip connection layer (Fig. 1) before inputting to the CAB, it contains noise pollution from the background areas; the 2nd channel of refined feature maps from the output of CAB in the decoder stage (Fig. 1). The background noise is reduced and the foreground vessel feature is highlighted via the channel attention operation.

where *GAP* means the operation of global average pooling, φ denotes 1 × 1 convolution followed by rectified linear unit and ϕ indicates 1 × 1 convolution followed by sigmoid activation. An intuitive display of CAB is shown at the left bottom of Fig. 1.

2.6. Data augmentation

As there are limited manually annotated datasets, data augmentation is necessary for the benefit of improving the model generalization. To teach SVS-net how to accommodate to various sample transformations, we adopt multiple augmentation methods including rotations by the angle ranging in $\mathbb{T}-10^{\circ}, 10^{\circ}\mathbb{U}$, flipping both horizontally and vertically, scaling by a factor of 0.2, random crop, affine transformations. For the images in our dataset, there is a 50% probability to perform each of above transformations to generate new samples in real time during the training process.

2.7. Loss function

To tackle the class imbalance problem in vessel segmentation, we employed Dice loss function to guide parameters learning. The class imbalance problem mainly has two aspects: firstly, the number of negative pixels (being 0, *i.e.*, background) is much more than the number of positive pixels (being 1, *i.e.*, vessel pixels); secondly, the ratio between the two classes varies a lot among both inter-frame in the same XCA sequence and intra-frame or inter-frame in different XCA sequences. Currently most semantic segmentation tasks adopted the following cross entropy (CE) (Mosinska, Marquez-Neila, Kozinski, & Fua, 2018; Ronneberger et al., 2015) to optimize the network:

$$L_{CE} \mathrm{D} \sum_{1}^{N} y_i \log p_i \mathrm{C} (1 - y_i) \log(1 - p_i) \tag{3}$$

It can be observed that from Eq. (3), each pixel contributes equally to the CE loss. Hence, CE loss tends to bias the network's optimization.

Different from CE loss calculated in pixel-wise form, Dice loss can avoid above problem by measuring the overlap ratio between ground truth mask and the predicted vessel mask. Dice loss is defined in Drozdzal et al. (2018), Zhang et al. (2018) as follows:

$$L_{DiceCoef} \mathrm{D} -\frac{2 \sum_{1}^{N} p_i y_i \mathrm{C} \epsilon}{\sum_{1}^{N} p_i \mathrm{C} \sum_{1}^{N} y_i \mathrm{C} \epsilon} \tag{4}$$

where y_i 2 f0, 1g is ground truth label and p_i 2 T0, 1U is predicted value for location *i*. *N* is the total number of pixels, ϵ is a very small constant used to keep value stable. From Eq. (4) we can find that the Dice loss is applied to the whole mask and it measures the overall loss for that mask rather than the average loss across all the pixels.

3. Experiment results

3.1. Materials

In our experiments, 120 sequences of real clinical X-ray coronary angiograms images are acquired from Renji Hospital of Shanghai Jiao Tong University. The length of each sequence ranges from 30 to 140 frames. Images from 120 sequences have been manually annotated by three experts to constitute the ground truth. Specifically, for the totally hard-annotated 323 samples from these 120 annotated sequences including extremely low-contrast vessels and vessel trees that contain a lot of thin vessel branches, we take three experts' average annotated result as the final ground truth.

It is worth noting that these XCA sequences in the dataset are acquired from different machines (*i.e.*, 800 mAh digital silhouette angiography X-ray machine from Siemens, medical angiography

Fig. 5. The loss curve and its local enlarged curve for both training set and validation set in training process.

X-ray system from Philips), the resolution, the noise distribution and the pixels' intensity range of each single frame are different. To eliminate these differences, we resize the images from the XCA sequence into 512 × 512 resolution with 8 bits per pixel, employ Poisson denoising methods (Cerciello, Bifulco, Cesarelli, & Fratini, 2012) to smooth the noise and normalize the pixels' intensity range into 0 − 1.

Furthermore, due to the varieties of XCA images with different directions and angles of X-ray penetration as well as different patient sources with different dosages of contrast agents, the vessels visibility of different sequences in clinic is quite changeable. Thus, designing a robust vessel segmentation algorithm is necessary for the XCA data with poor visual quality. Besides, proper selection of frames from each sequence for experiment is crucial (Lim & Keles, 2018) especially when both of the background and foreground are dynamic and contain many artifacts. The strategy of selecting the training frames is similar to Wang, Luo and Jodoin (2017). We selected XCA images containing most of vessel structures as experiment samples from 120 annotated sequence according to their lengths and visual quality. Totally, 332 samples are obtained for our experiment. The dataset is randomly divided into training dataset, validation dataset, and test data at approximately 0.5, 0.25 and 0.25, respectively.

We investigate the proposed model's performance on the abovementioned dataset. We plot the loss curves for both the training set and validation set in the training process. As can be seen in Fig. 5, for both training set and validation set, the loss reduces quickly at the beginning stage of training process, and gradually converges. There is no sign that the model falls into over-fitting or under-fitting state. Meanwhile, the size of our dataset is assumed to be properly matched into the size of our model.

All the experiments performed in this work were approved by our institutional review board. The dataset which will be released to public has received the transfer agreement from our cooperative partners. All the dataset is stored in mat array format according to the corresponding filenames, and they will be available on website.[3] You can also visit the website to access further detailed information on the dataset.

[3] The source codes and dataset will be available at https://github.com/Binjie-Qin/SVS-net.

3.2. Evaluation metrics

Several metrics, namely, detection rate (DR), precision (P), and F measure are employed to quantitatively evaluate the performances of our segmentation method and also compare them with other state-of-the-art methods. The above metrics are defined as below:

$$DR = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP}, \quad F = \frac{2 \quad DR \quad P}{DR + P} \tag{5}$$

where TP (true positives) is the total number of correctly classified pixels in vessel regions of the predicted vessel probability map, FP (false positives) indicates the total number of wrongly identified as vessel pixels but practically belonging to backgrounds in the predicted vessel probability map, TN (true negatives) and FN (false negatives) represent the total number of correctly predicted as background pixels and wrongly predicted as background pixels in the predicted output, respectively. DR measures the proportion between the correctly identified vessel pixels and the total vessel pixels in the ground truth, P measures the ratio of true positives among all the true positives. F measure comprehensively considers both P and DR metrics and indicates the overall segmentation performance. All these metrics range in [0, 1], and a higher value indicates better segmentation performance.

3.3. 2D vs 3D with and without channel attention mechanism

We utilize 3D convolution layers to extract rich temporal–spatial feature representation in encoder network. To investigate whether the temporal–spatial features are more advanced compared to purely spatial features extracted by 2D convolutional layers for generating final predicted probability map, we replace 3D convolutional layers with corresponding 2D ones in encoder network while keeping the decoder network the same. It is noted that simple substitution reduces the number of trainable parameters and hence weakens the model's expressive capacity. For fair comparison, we increase the number of convolution layers in the encoder network for 2D version to make both 3D version and 2D version have comparable amount of parameters. In addition, we investigate the effectiveness of CAB by removing it from decoder

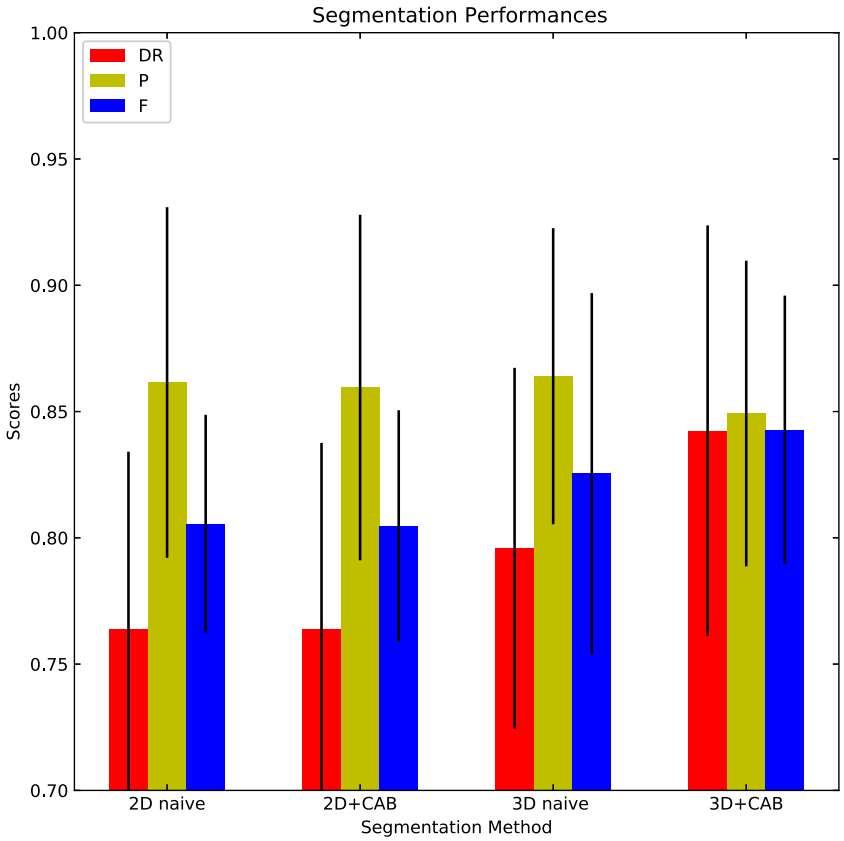

Fig. 6. Vessel segmentation performance using 2D and 3D convolutional layers with and without CAB. The detection rate (DR), precision (P) and F measure of test data.

Table 2
The average detection rate (DR), precision (P) and F measure (mean value ± standard deviation) for test data.

Method	DR	P	F
2D naive	0.7640 ± 0.0701	0.8615 ± 0.0694	0.8056 ± 0.0431
2D+CAB	0.7638 ± 0.0738	0.8595 ± 0.0684	0.8046 ± 0.0459
3D naive	0.7959 ± 0.0714	0.8640 ± 0.0586	0.8255 ± 0.0714
3D+CAB	**0.8424 ± 0.0813**	0.8492 ± 0.0605	**0.8428 ± 0.0531**

network in SVS-net and 2D model respectively. We choose our 2D model without CAB (2D naive) as baseline, and compare it with 2D model with CAB (2D+CAB), 3D model without CAB (3D naive) and 3D model with CAB (3D+CAB) respectively.

To quantitatively evaluate the performance, we measure three metrics on the test set. The results are shown in Table 2 and Fig. 6. Specifically, we compare the performance between different feature extraction manners and then analyze the function of CAB. As can be seen in Table 2, compared with the baseline, the 3D naive model obtains higher scores in terms of DR (79.59%), P (86.40%), F (82.55%) measures and surpasses its 2D counterpart by 4.17%, 0.29% and 2.47%, respectively. Despite they have almost similar model complexity, their performance have big differences. This is because 3D version integrating contextual spatial–temporal features while 2D version only using spatial domain features to predict the final probability map. Obviously, the former one has more sufficient and robust information in discriminating between noisy artifacts and vessel trees.

Next, we investigate the effect of CAB in both 3D and 2D scenarios. From Table 2, we can find that the channel attention strategy decreases the P measures from 0.8615 to 0.8595 and reduces the DR scores from 0.7640 to 0.7838 in 2D case. There are slight changes in both DR and P. Hence, F measures almost keep consistent due to F measure achieving the trade-off between P (related with FPs) and DR (related with FNs) measures. In 3D case, the CAB also shows the good compromise between P and DR (*i.e.*, DR increases from 0.7959 to 0.8424 while P declines from 0.8640 to 0.8492) but the overall performance F measure improves by 2.09% and arrives at 0.8428. We analyze the role that the CAB plays in the overall performance: in the 2D case, we find the fact that CAB makes a compromise between P and DR with rare improvement in F measure, which is now interpreted that 2D version failing to provide sufficient and valuable information for CAB to choose from; however, in 3D case, the temporal–spatial information is relatively rich so that the CAB can suppress trivial features in noisy background and utilize the most discriminative ones in foreground to generate fine vessel mask.

Furthermore, we analyze the stability of SVS-net's segmentation performance. When we compare SVS-net (3D+CAB) with the baseline (2D naive), the DR and F witness a large increase by 10.26% and 4.61%, respectively. The metrics have a relatively high standard deviation, which indicates that there exist both relatively hard-segmented and relatively easy-segmented samples in test sets. The relatively hard-segmented samples may pull down the whole metrics and result in the high standard deviation. In the future, we will enlarge our dataset and increase the number of hard-segmented samples to help SVS-net pay more attention to them for improving the its performance on those samples. It is worth noting that the relatively high standard deviation problem almost exists in all the tested methods, which illustrate there indeed exist hard-segmented samples. The comparative methods also fail to deal with these hard-segmented samples. In view of the average metrics, we further our confidence that SVS-net performs better than all the other methods.

Moreover, the 3D+CAB model achieves the highest DR score while the lower P score. Higher DR score means lower FNs, which

Fig. 7. Four instances of vessel segmentation result by different vessel segmentation methods. From left to right, each row displays the original XCA image, the manually outlined ground truth vessel segmentation, the vessel images segmented by 2D naive, 2D+CAB, 3D naive, 3D+CAB, respectively.

Table 3
The average detection rate (DR), precision (P) and F measure (mean value standard deviation) for test data.

Method	DR		P		F	
CE loss	0.8197	0.0814	0.8423	0.0633	0.8262	0.0428
Dice loss	0.8424	0.0813	0.8492	0.0605	0.8428	0.0531

indicates SVS-net have superior capacity in detecting vessel pixels from backgrounds. In detecting vessel pixels, it is likely to mistakenly identify backgrounds that resemble vessel pixels as vessel pixels. This may increase FPs to some degree, so that the P score will suffer from degradation. However, the 3D+CAB model's overall F measure performance is highest, which illustrates its balanced and better performance in accurately recognizing both vessel pixels and background pixels when compared with other methods.

Intuitive quantization result can be seen in Fig. 6. The typical segmentation results are displayed in Fig. 7. As shown in Fig. 7, the vessel masks produced by SVS-net have less fractures (i.e., less FNs) than do the other methods, which implies SVS-net's better performance on detecting vessel pixel (TPs). In 2D settings, the vessel masks either have more fracture (i.e., more FNs) or have more artifacts (i.e., more FPs). This phenomenon shows that their pool ability on differentiating vessel pixels from background pixels. Through above observation and analysis, we validate the effectiveness of the 3D convolutional layers and CAB used in SVS-net.

3.4. Cross entropy loss vs Dice loss

To validate the loss function in deep network for vessel segmentation, we employ CE loss and Dice loss function as objective function to train our model respectively and use the same settings with other parts of network. It can be seen in Table 3, the model trained with Dice loss can result in the performance improvement by 2.0% in terms of F measure, which implies the effectiveness of Dice loss in class imbalance segmentation task when compared with the CE loss. Besides, the model trained with Dice loss achieves higher F measure and DR score but slightly higher P score when compared with the model trained with CE loss. The Dice loss's optimization goal is expected to facilitate the model in getting higher gain of F measure for achieving better overall segmentation performance. When the model is optimized, the overall segmentation performance of F measure certainly has an upper bound. With this upper bound on the F measure that is computed from the harmonic mean of DR and P, the model makes a trade-off balance between the DR and P metrics, so that it may result in the balance with higher DR and slightly higher P.

Then, we compare the intuitive segmentation results by two loss functions. When compared with the vessel masks obtained from the model trained with Dice loss, the vessel masks obtained from the CE loss have blurred boundaries (see the red arrows in Fig. 8(c) (d)), which means pixels at vessel boundaries are less confident to discriminate whether they belong to vessels or backgrounds. Therefore, to get binary vessel masks, we should carefully apply proper threshold to the original probability maps which involves troublesome manual operations (shown in Fig. 8(e) (h)). While the masks produced by model trained with Dice loss have clear boundaries, there is no need to utilize threshold anymore. Therefore, the Dice loss function is appropriate to train SVS-net for sequential vessel segmentation.

3.5. Comparison with other state-of-the-art methods

We compare SVS-net with three traditional vessel segmentation algorithms, i.e., Coye's filter method (Coye's) [4] (Coye, 2017), Jin's spatially adaptively filtering method (Jin's) (Jin et al., 2018), Kerkeni's multi-scale region growing method (Kerkeni's) (Kerkeni et al., 2016), and four deep learning-based methods, i.e., Retinalnet,[5] (Liskowski & Krawiec, 2016; Ronneberger et al., 2015)

4 http://www.mathworks.com/matlabcentral/fileexchange/50839.
5 https://github.com/orobix/retina-unet.

Table 4
The average detection rate (DR), precision (P) and F measure (mean value standard deviation) for test data by state-of-the-art methods and our method.

Method	DR		P		F		param	Inference time
Coye's	0.5694	0.3096	0.2127	0.1365	0.2963	0.1752		0.071s
Jin's	0.6127	0.1948	0.7715	0.2126	0.6639	0.1677		11.61s
Kerkeni's	0.6703	0.1322	0.7348	0.1321	0.6863	0.1047		4.708s
Retinal-net	0.7708	0.1003	0.6807	0.1160	0.7141	0.0865	0.47M	2.28s
SU-Net	0.6914	0.1057	0.9018	0.0785	0.7734	0.0580	15.32M	0.046s
X-ray net	0.7974	0.0748	0.7780	0.1038	0.7794	0.0586	14.1M	0.053s
BTS-DSN	0.7251	0.0971	0.8626	0.0750	0.7803	0.0612	7.8M	0.067s
SVS-net	0.8424	0.0813	0.8492	0.0605	0.8428	0.0531	10.2M	0.178s

Fig. 8. Original vessel segmentation result by Dice loss, CE loss and threshold (0.2, 0.4, 0.6, 0.8 respectively) postprocessing for the segmentation result by CE loss. (a)

Fig. 10. Four instances of vessel segmentation for real XCA image sequence by different vessel segmentation methods. From left to right, each row displays the original XCA image, the manually outlined ground truth vessel segmentation, the vessel images segmented by Coye's, Jin's, Kerkeni's, Retinal-net, SU-Net, X-ray net, BTS-DSN, and SVS-net, respectively.

temporally consistent information. It not only increases temporal information but also introduces disturbances at the same time. BTS-DSN adopts deeply supervised strategy and achieves relative higher metrics. However, there are still FPs in the vessel regions. Compared with above deep network methods, SVS-net can not only robustly detect the vessel regions with almost intact vessel structures with continuous vessel branches but also effectively remove the noisy background artifacts. The continuity and integrity of the segmented vessel branches is assumed to be owed to the contextual information inferred in the temporal spatial features extracted by the encoder network and feature fusion in the skip connection layers. The noise reduction in the segmented vessel regions is mostly derived from the discriminative feature selection implemented by the channel attention mechanism. Therefore, the temporal spatial feature extraction, feature fusion and the discriminative feature learning adopted in SVS-net are necessary to help improve the segmentation performances.

Additionally, there is a small number of thin vessel branches fail to be recognized by SVS-net. It is really challenging and we plan to design novel loss function in our future work, which will differentiate the thick and thin vessels efficiently and integrate these different vessels with different weights. In this way, we increase the weights of thin vessels in loss function and promote the model to pay more attention to the thin vessels. Thin-vessel segmentation is definitely a promising direction for improving the clinical value of XCA images.

Our experiments are implemented on GPU (i.e., NVIDIA 1080 Ti, 11 GB). The number of parameters and the average runtime of per test image for deep-learning methods are listed in Table 4. Compared with other deep learning-based methods having bigger or fewer number of parameters, i.e., from the 15.32 million parameters for SU-Net to the 0.47 million parameters for Retinal-net, as well as having longer or shorter inference time, i.e., from the 2.28 s for Retinal-net to the 0.046 s for SU-Net, SVS-net has 10.2 million parameters and 0.178 s inference times to achieve an intermediate level of complexity. The reason for SVS-net's medium-complexity in achieving its best segmentation performance is two-fold: (1) the 3D convolutional layers instead of 2D convolutional layers are adopted in the stage of feature extraction; (2) the fully connection layers are utilized in the stage of feature refinement. Although these two strategies for feature extraction and feature refinement employed in SVS-net explicitly increase the parameter number, they are necessary as the verification in the hyper-parameter experiments in Section 2.2. Moreover, the relatively long inference time mainly results from the feature fusion and channel attention mechanism. In the future work, we intend to explore more efficient network architectures for further decreasing computation time and improving inference efficiency.

3.6. Downstream works

Vessel segmentation is an efficient preprocessing procedure for various medical tasks. To assess the influence of vessel segmentation on various medical tasks, we further investigate two down-stream tasks that use the proposed SVS-net. We choose three state-of-the-art segmentation methods (i.e., SU-Net, BTS-DSN, X-ray net) to compare our SVS-net.

Estimating the distribution of coronary vessel networks via vessel segmentation is very important to evaluate the coronary circulation (Vigneshwaran, Sands, LeGrice, Smaill, & Smith, 2019) in percutaneous coronary intervention. Usually, we estimate the area proportion of vessel network distribution in the whole heart regions of XCA images. Obviously, the wider the distribution, the smoother the blood flows in coronary circulation. Specifically, we use relative gland volume errors (GVEs) defined in Nooshin et al. (2019) to measure the vessel distribution area. GVE is calculated by the absolute difference between the predicting segmentation $V\ y_{pr}$ and the manual ground-truth segmentation $V\ y_{gt}$:

$$GVE D \frac{V\ y_{gt} \quad V\ y_{pr}}{V\ y_{gt}} \quad 100\% \tag{6}$$

$V\ .\ /$ is based on counting the positive voxels in the binary segmentation. From the definition, it is easy to learn that a good segmentation method should have low GVE value. The relative GVE is summarized in Table 5. From Table 5, we can see that the mean GVE of SVS-net is 9:74%, which is much lower than those of other methods. Besides, the standard deviation of our method is also lower than those of other methods. These measures implicate that SVS-net is better and more stable in vessel distribution estimation than other methods.

Furthermore, quantitative coronary analysis and corresponding myocardial perfusion analysis are other downstream works for the diagnosis and therapy of coronary artery disease. The gray-level intensities of cardiac vessels carry important information for the quantitative analysis. Usually, we locate the vessel through

Table 5
Vessel network volume calculations between the manual ground-truth and segmentation of each method.

Method	X-ray net	SU-Net	BTS-DSN	SVS-net
Relative GVE difference (%) mean ± std	13.00 ± 10.35	23.25 ± 12.30	17.41 ± 11.47	**9.74** ± **7.52**

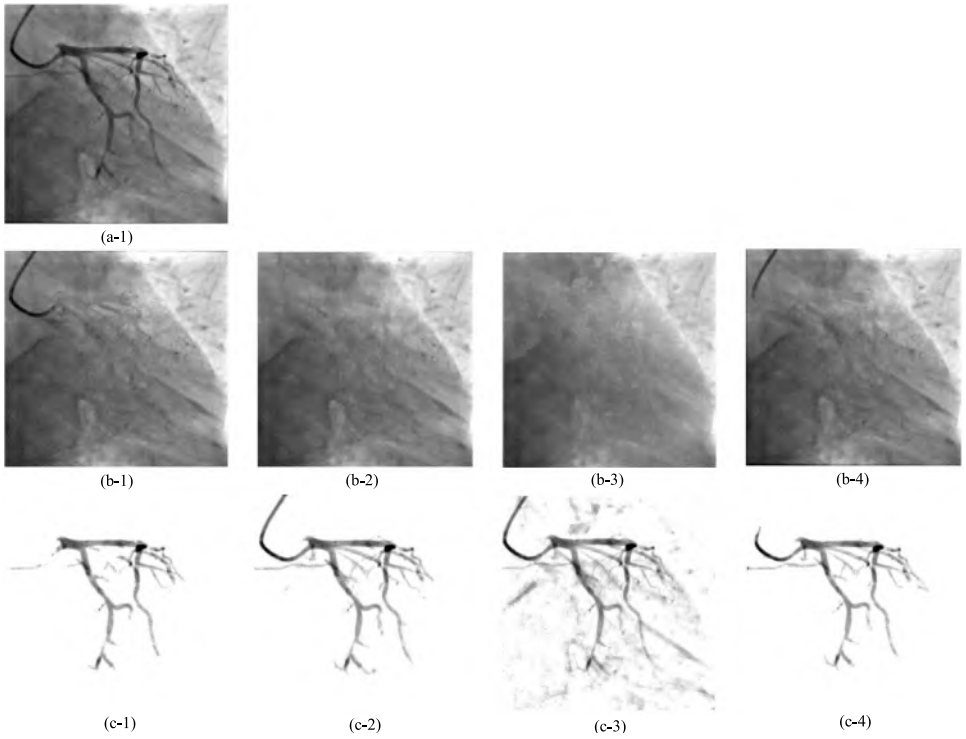

Fig. 11. Vessel gray intensity recovery. From top to bottom, each row displays the original XCA image, the background and foreground vessel images recovered from SU-Net, X-ray net, BTS-DSN and SVS-net, respectively.

segmentation but lose the gray intensity at the same time. As illustrated in Qin et al. (2019), we can use the vessel mask regions from the vessel segmentation and then complete vessel gray information in these regions by tensor completion algorithm (Qin et al., 2019). We compare three segmentation methods' effect on the final gray intensity recovery after vessel mask region segmentation. As shown in Fig. 11, SVS-net is much more conducive to reconstruct vessels and their gray intensities from the complex and noisy backgrounds of X-ray images. There are few vessel residuals remained in the background regions. Hence, SVS-net can provide relatively accurate vessel mask segmentation for recovering gray intensity in quantitative coronary analysis.

4. Discussion and conclusion

We propose a sequential vessel segmentation deep network, which integrates 3D convolutional layers extracting rich temporal–spatial features and utilizes CAB learning discriminative features from the complex and noisy background artifacts in the XCA image sequences. Experiment results verify the superior performance of special designs in our SVS-net. The proposed SVS-net can effectively segment the whole branches of the vessel trees from the XCA sequences. There is still room in the future work to improve the accuracy of segmenting small branches of thin vessels and enhance temporal–spatial consistency of vessel tree mask. To achieve a reliable segmentation of all small and thin vessels in the low-contrast and noisy XCA sequence, the channel-wise attention scheme can be further integrated into saliency-aware image matching (Qin et al., 2018; Qin, Shen, Zhou, Zhou, & Lv, 2016) and segmentation (Wang, Shen, Yang, & Porikli, 2018) methods as well as pixel-wise phase-based edge feature filtering (Mei, Hu, Fei, & Qin, 2020; Reisenhofer & King, 2019; Zhao, Zheng, Liu, Zhao, Luo, Yang, Na, Wang, & Liu, 2018) to automatically choose the key frame and corresponding regions that contain the most salient small and thin vessel features from the XCA sequence so that this frame's thin vessel feature representation can be taken as priors for pixel-wise labeling in sequential vessel segmentation. Deep feature matching (Kong, Supancic, Ramanan, & Fowlkes, 2019) and deep temporal–spatial correlation (Wang, Luo, Shen, & Pantic, 2019) in the image sequence with deep autoencoding prior (Li, Qin, Xiao, Liu, Wang, & Liang, 2020; Zhang, Zhang, Qin, Zhang, Xu, Liang, & Liu, 2020) can also be utilized to transfer the learning priors from key frame to its neighboring frames containing unsharp small vessels.

Furthermore, we can sample vessel mask regions using contrast agent motion information or randomly sample vessel mask regions for background inpainting via tensor (or matrix) completion (Qin et al., 2019; Unberath et al., 2017) and deep video inpainting (Kim, Woo, Lee, & Kweon, 2020), the completed background is then subtracted from XCA image sequence for the overall vessel extraction. This scheme of trial-and-completion can not only accurately recover the structures and intensities of vessel trees but also well compensate the deficiency of small vessel extraction (or segmentation) in the XCA image sequences. Such vessel extraction can be effectively implemented in an unsupervised deep network (Sultana, Mahmood, Javed, & Jung, 2019).

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.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61271320, 81370041 and 81400261) and Shanghai Jiao Tong University Cross Research Fund for Translational Medicine , China (ZH2018ZDA19). Yueqi Zhu was partially supported by three-year plan program by Shanghai Shen Kang Hospital Development Center, China (16CR3043A). BF was partially supported by NIH, USA Grants R01CA156775, R21CA176684, R01CA204254, and R01HL140325. The authors would like to thank all authors for opening source codes used in the experimental comparison in this work. The authors would also like to thank the anonymous reviewers whose contributions considerably improved the quality of this paper.

References

Albarqouni, S., Fotouhi, J., & Navab, N. (2017). X-ray in-depth decomposition: revealing the latent structures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 444–452). Springer.

Ambrosini, P., Ruijters, D., Niessen, W. J., Moelker, A., & van Walsum, T. (2017). Fully automatic and real-time catheter segmentation in x-ray fluoroscopy. In *International conference on medical image computing and computer-assisted intervention* (pp. 577–585). Springer.

Blondel, C., Malandain, G., Vaillant, R., & Ayache, N. (2006). Reconstruction of coronary arteries from a single rotational X-ray projection sequence. *IEEE Transactions on Medical Imaging*, *25*(5), 653–663.

Carrillo, J. F., Hoyos, M. H., Dávila, E. E., & Orkisz, M. (2007). Recursive tracking of vascular tree axes in 3D medical images. *International Journal of Computer Assisted Radiology and Surgery*, *1*(6), 331–339.

Cerciello, T., Bifulco, P., Cesarelli, M., & Fratini, A. (2012). A comparison of denoising methods for X-ray fluoroscopic images. *Biomedical Signal Processing and Control*, *7*(6), 550–559.

Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., & Goldbaum, M. (1989). Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Transactions on Medical Imaging*, *8*(3), 263–269.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, D., Zhang, J., & Cohen, L. D. (2019). Minimal paths for tubular structure segmentation with coherence penalty and adaptive anisotropy. *IEEE Transactions on Image Processing*, *28*(3), 1271–1284.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., et al. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5659–5667).

Coye, T. (2017). A novel retinal blood vessel segmentation algorithm for fundus images. *MATLAB Central File Exchange*.

Dasgupta, A., & Singh, S. (2017). A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation. In *Biomedical imaging (ISBI 2017), 2017 IEEE 14th international symposium on* (pp. 248–251).

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350.

Dehkordi, M. T., Hoseini, A. M. D., Sadri, S., & Soltanianzadeh, H. (2014). Local feature fitting active contour for segmenting vessels in angiograms. *IET Computer Vision*, *8*(3), 161–170.

Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Di Jorio, L., Tang, A., et al. (2018). Learning normalized inputs for iterative estimation in medical image segmentation. *Medical Image Analysis*, *44*, 1–13.

Fan, J., Yang, J., Wang, Y., Yang, S., Ai, D., Huang, Y., et al. (2018). Multichannel fully convolutional network for coronary artery segmentation in X-ray angiograms. *IEEE Access*, *6*, 44635–44643.

Felfelian, B., Fazlali, H. R., Karimi, N., Soroushmehr, S. M. R., Samavi, S., Nallamothu, B., et al. (2016). Vessel segmentation in low contrast X-ray angiogram images. In *2016 IEEE international conference on image processing* (pp. 375–379). IEEE.

Frangi, A. F., Niessen, W. J., Vincken, K. L., & Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *International conference on medical image computing and computer-assisted intervention* (pp. 130–137). Springer.

Fu, H., Xu, Y., Wong, D. W. K., & Liu, J. (2016). Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In *2016 IEEE 13th international symposium on biomedical imaging* (pp. 698–701).

Fu, K., Zhao, Q., & Gu, I. Y.-H. (2018). Refinet: A deep segmentation assisted refinement network for salient object detection. *IEEE Transactions on Multimedia*, *21*(2), 457–469.

Ghavami, N., Hu, Y., Gibson, E., Bonmati, E., Emberton, M., Moore, C. M., et al. (2019). Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Medical Image Analysis*, *58*, 101558.

Guo, S., Gao, Y., Wang, K., & Li, T. (2018). Deeply supervised neural network with short connections for retinal vessel segmentation. ArXiv preprint arXiv:1803.03963.

Hao, H., Ma, H., & van Walsum, T. (2018). Vessel layer separation in x-ray angiograms with fully convolutional network. In *Medical imaging 2018: Image-guided procedures, robotic interventions, and modeling (Vol. 10576)* (p. 105761V). International Society for Optics and Photonics.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hu, J., Wang, H., Gao, S., Bao, M., Liu, T., Wang, Y., et al. (2019). S-unet: A bridge-style u-net framework with a saliency mechanism for retinal vessel segmentation. *IEEE Access*, *7*, 174167–174177. http://dx.doi.org/10.1109/access.2019.2940476.

Hu, K., Zhang, Z., Niu, X., Zhang, Y., Cao, C., Xiao, F., et al. (2018). Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. *Neurocomputing*, *309*, 179–191.

Huang, C., Change Loy, C., & Tang, X. (2016). Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5175–5184).

Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. (2018). Learn to pay attention. ArXiv preprint arXiv:1804.02391.

Jin, M., Hao, D., Ding, S., & Qin, B. (2018). Low-rank and sparse decomposition with spatially adaptive filtering for sequential segmentation of 2D+ t vessels. *Physics in Medicine and Biology*, *63*(17), 17LT01.

Jin, M., Li, R., Jiang, J., & Qin, B. (2017). Extracting contrast-filled vessels in X-ray angiography by graduated RPCA with motion coherency constraint. *Pattern Recognition*, *63*, 653–666.

Makowski, P., Sørensen, T. S., Therkildsen, S. V., Materka, A., Stødkilde-Jørgensen, H., & Pedersen, E. M. (2002). Two-phase active contour method for semiautomatic segmentation of the heart and blood vessels from MRI images for 3D visualization. *Computerized Medical Imaging and Graphics*, *26*(1), 9–17.

Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., & Van Gool, L. (2016). Deep retinal image understanding. In *International conference on medical image computing and computer-assisted intervention* (pp. 140–148). Springer.

Manniesing, R., Viergever, M. A., & Niessen, W. J. (2007). Vessel axis tracking using topology constrained surface evolution. *IEEE Transactions on Medical Imaging*, *26*(3), 309–316.

Mei, K., Hu, B., Fei, B., & Qin, B. (2020). Phase asymmetry ultrasound despeckling with fractional anisotropic diffusion and total variation. *IEEE Transactions on Image Processing*, *29*, 2845–2859.

Mo, J., & Zhang, L. (2017). Multi-level deep supervised networks for retinal vessel segmentation. *International Journal of Computer Assisted Radiology and Surgery*, *12*(12), 2181–2193.

Moccia, S., Momi, E. D., Hadji, S. E., & Mattos, L. S. (2018). Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. *Computer Methods and Programs in Biomedicine*, *158*, 71–91.

Mosinska, A., Marquez-Neila, P., Kozinski, M., & Fua, P. (2018). Beyond the pixel-wise loss for topology-aware delineation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 3136–3145).

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S. R., Ward, K., Jafari, M. H., et al. (2016). Vessel extraction in x-ray angiograms using deep learning. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society* (pp. 643–646). IEEE.

Peng, Y., He, X., & Zhao, J. (2018). Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, *27*(3), 1487–1500.

Pfeuffer, A., & Dietmayer, K. (2019). Separable convolutional LSTMs for faster video segmentation. ArXiv preprint arXiv:1907.06876.

Pfeuffer, A., Schulz, K., & Dietmayer, K. (2019). Semantic segmentation of video sequences with convolutional LSTMs. ArXiv preprint arXiv:1905.01058.

Qin, B., Jin, M., Hao, D., Lv, Y., Liu, Q., Zhu, Y., et al. (2019). Accurate vessel extraction via tensor completion of background layer in X-ray coronary angiograms. *Pattern Recognition*, *87*, 38–54.

Qin, B., Shen, Z., Fu, Z., Zhou, Z., Lv, Y., & Bao, J. (2018). Joint-saliency structure adaptive kernel regression with adaptive-scale kernels for deformable registration of challenging images. *IEEE Access*, *6*, 330–343.

Qin, B., Shen, Z., Zhou, Z., Zhou, J., & Lv, Y. (2016). Structure matching driven by joint-saliency-structure adaptive kernel regression. *Applied Soft Computing*, *46*, 851–867.

Rashed, H., Yogamani, S., El-Sallab, A., Krizek, P., & El-Helw, M. (2019). Optical flow augmented semantic segmentation networks for automated driving. ArXiv preprint arXiv:1901.07355.

Reisenhofer, R., & King, E. J. (2019). Edge, ridge, and blob detection with symmetric molecules. *SIAM Journal on Imaging Sciences*, *12*(4), 1585–1626.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Ruder, M., Dosovitskiy, A., & Brox, T. (2018). Artistic style transfer for videos and spherical images. *International Journal of Computer Vision*, *126*(11), 1199–1219.

Sakkos, D., Ho, E. S., & Shum, H. P. (2019). Illumination-aware multi-task GANs for foreground segmentation. *IEEE Access*, *7*, 10976–10986.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, *53*, 197–207.

Simoncelli, E. P., Adelson, E. H., & Heeger, D. J. (1991). Probability distributions of optical flow. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition* (pp. 310–315). IEEE.

Soares, J. V., Leandro, J. J., Cesar, R. M., Jelinek, H. F., & Cree, M. J. (2006). Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, *25*(9), 1214–1222.

Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., & Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, *23*(4), 501–509.

Sultana, M., Mahmood, A., Javed, S., & Jung, S. K. (2019). Unsupervised deep context prediction for background estimation and foreground segmentation. *Machine Vision and Applications*, *30*(3), 375–395.

Sun, D., Yang, X., Liu, M.-Y., & Kautz, J. (2018). Pwc-net: cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8934–8943).

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (pp. 4278–4284).

Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., & Nichols, M. (2016). Cardiovascular disease in europe: epidemiological update 2016. *European Heart Journal*, *37*(42), 3232–3245.

Unberath, M., Aichert, A., Achenbach, S., & Maier, A. (2017). Consistency-based respiratory motion estimation in rotational angiography. *Medical Physics*, *44*(9), e113–e124.

Vigneshwaran, V., Sands, G. B., LeGrice, I. J., Smaill, B. H., & Smith, N. P. (2019).