Computers in Biology and Medicine 118 (2020) 103657

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint

Wei Wu [a,b], Jingyang Zhang [a,b], Hongzhi Xie [c,**], Yu Zhao [a,b], Shuyang Zhang [c], Lixu Gu [a,b,*]

[a] Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] Instituted of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[c] Department of Cardiology, Peking Union Medical College Hospital, Peking, 100005, China

ARTICLE INFO

Keywords:
X-ray coronary angiography
Coronary artery stenosis detection
Convolutional neural network
Temporal constraint

ABSTRACT

Coronary artery disease (CAD) is a major threat to human health. In clinical practice, X-ray coronary angiography remains the gold standard for CAD diagnosis, where the detection of stenosis is a crucial step. However, detection is challenging due to the low contrast between vessels and surrounding tissues as well as the complex overlap of background structures with inhomogeneous intensities. To achieve automatic and accurate stenosis detection, we propose a convolutional neural network-based method with a novel temporal constraint across X-ray angiographic sequences. Specifically, we develop a deconvolutional single-shot multibox detector for candidate detection on contrast-filled X-ray frames selected by U-Net. Based on these static frames, the detector demonstrates high sensitivity for stenoses yet unacceptable false positives still exist. To solve this problem, we propose a customized seq-fps module that exploits the temporal consistency of consecutive frames to reduce the number of false positives. Experiments are conducted with 148 X-ray angiographic sequences. The results show that the proposed method outperforms existing stenosis detection methods, achieving the highest sensitivity of 87.2% and positive predictive value of 79.5%. Furthermore, this study provides a promising tool to improve CAD diagnosis in clinical practice.

1. Introduction

Coronary artery disease is the most common type of heart disease and a major cause of mortality worldwide [1]. It occurs when obstructive atherosclerotic plaque builds up in the inner walls of coronary arteries. This causes stenosis, i.e., the narrowing or occlusion of the coronary artery lumen, leading to severe symptoms such as angina and even myocardial infarction. X-ray coronary angiography (XCA) is currently regarded as the gold standard for coronary artery stenosis detection. With an injected contrast agent, XCA can offer anatomical information of even very small vessels and enable cardiologists to observe dynamically from different projection angles. Cardiologists can then identify and locate each stenosis with a visual assessment. Fig. 1 shows XCA frames with annotated stenoses.

Manual detection of stenosis is subjective and time-consuming, requiring rich clinical experience. Therefore, developing an XCA-based automatic detection algorithm can improve diagnostic efficiency and confidence. However, it is also challenging due to complex vessel structures, poor contrast between vessels and surrounding tissues, nonuniform illumination, and overlap of background structures with inhomogeneous intensities. Existing methods are mainly based on computed tomography angiography (CTA), which can be divided into three main categories: (1) lumen segmentation-based methods; (2) arterial wall segmentation-based methods; and (3) centerline extraction-based methods. Lumen segmentation-based methods identify stenosis by measuring the lumen diameter. Shahzad et al. [2] extracted the centerline and employed graph cuts and robust kernel regression to segment arterial lumens. Stenosis was identified by comparing the real diameter of the segmented lumen with the expected diameter of the modeled healthy lumen. Arterial wall segmentation-based methods identify stenosis by measuring the diameter difference between the inner and outer walls. Wang et al. [3] adopted a level-set model to segment the inner and outer arterial walls. A large difference between the two diameters indicated the existence of stenosis. Similarly, Broersen et al. [4] first detected arterial wall contours and then used a regression model to calculate the deviations from normal vessels to

* Corresponding author. Instituted of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China.
** Corresponding author.
E-mail addresses: xiehongzhi@medmail.com.cn (H. Xie), gulixu@sjtu.edu.cn (L. Gu).

https://doi.org/10.1016/j.compbiomed.2020.103657
Received 29 October 2019; Received in revised form 8 February 2020; Accepted 8 February 2020
Available online 12 February 2020

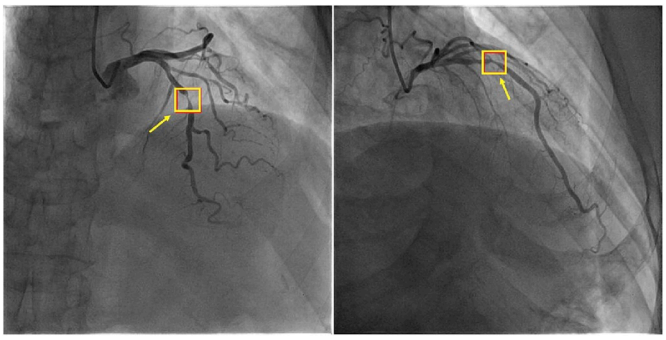


Fig. 1. XCA frames from two different sequences. The coronary arteries enhanced with contrast agent are the dark tubular structures. The stenoses are annotated with yellow rectangles and arrows.

detect stenosis. Centerline extraction-based methods identify stenosis by analyzing the image patches along the vessel centerline. Zerik et al. [5] extracted patches from multiplanar reformation CTA image volumes. A recurrent neural network was adopted to process the patch features along the centerline and determine the position of a stenosis. Despite these achievements, stenosis detection based on CTA images is generally influenced by low resolution, motion artifacts and severe vascular calcification. Therefore, XCA is still indispensable for cardiologists in clinical practice.

However, compared with CTA, relatively few studies are based on XCA. Related works can be divided into two main categories: semi-automatic methods and fully-automatic methods. Semiautomatic methods ignore the automatic detection task. They require human interactions to locate a stenosis and only focus on assessing its severity. After vessel structures were extracted with different methods, e.g., deformable spline with string matching [6] and Hessian vesselness filter with wavelet-based image fusion [7], the diameter of the target stenosis was manually measured to evaluate the severity. Fully-automatic methods concentrate on achieving automatic stenosis detection. The algorithm proposed by Wan et al. [8] integrated noise reduction, vascular structure enhancement, skeleton extraction, vessel diameter estimation and data postprocessing for stenosis detection. Similarly, Compas et al. [9] conducted vascular structure enhancement and extracted the skeleton for vessel diameter measurement. Stenosis identification was based on not only diameter variations but also spatio-temporal tracking. Despite a good performance, fully-automatic methods still have limitations: multiple preprocessing procedures, such as vessel enhancement, segmentation and skeleton extraction are performed in a single frame to detect stenosis. This is a time-consuming process, and the intrinsic complexities of XCA images prevent these substeps from obtaining ideal intermediate results. As a consequence, errors might accumulate, hampering final stenosis detection.

To avoid the cumbersome preprocessing steps for a single frame in traditional stenosis detection methods, we turn to convolutional neural network (CNN), the state-of-the-art method for object detection [10–14], to achieve stenosis detection. It is an end-to-end method with a strong feature extraction ability, generating detection results for every single frame quickly and directly. However, due to vessel motion and contrast agent flow, stenosis-like structures such as bent vessels or instantaneous contrast agent inhomogeneity appear in some of the frames, which might mislead the network to generate false positive detections. Therefore, detecting stenosis from single frames is not robust enough. Considering that these interferences are generally time-dependent, we propose to exploit the potential temporal characterization of the XCA sequence. Consecutive frames are selected with the corresponding CNN detection results summarized to remove false positives.

In conclusion, in this study, we propose a deep learning-based object detection network with temporal constraints on the XCA sequence to achieve automatic coronary artery stenosis detection. First, consecutive contrast-filled frames that are most beneficial for detection are selected by the segmentation network U-Net [15] to provide necessary temporal information. Then, a deconvolutional single-shot multibox detector (DSSD) [16] is applied to conduct stenosis detection directly on the selected raw X-ray angiograms without multiple preprocessing steps, providing static detection results for every single frame. Finally, with the designed temporal module called "sequence-false positive suppression" (seq-fps), we exploit the potential temporal consistency of the selected frames and produce constraints, which removes false positives and generates the final detection results.

The main contributions of this study are as follows: first, we propose a new framework for coronary artery stenosis detection with an XCA sequence. Second, to the best of our knowledge, this is the first work that has focused on coronary artery stenosis detection in X-ray angiograms using a deep learning-based object detection method. Third, we design the seq-fps module to exploit the potential temporal consistency of consecutive XCA frames, which is effective in false positive suppression. In general, the proposed method is superior in stenosis detection, outperforming traditional methods.

2. Materials and methods

2.1. Data description

The raw XCA sequence data used in this study were acquired from the Department of Cardiothoracic Surgery, Peking Union Medical College Hospital of China. They were collected from 63 patients (40 men and 23 women) with ages ranging from 51 to 67 years. The patients all underwent femoral artery cardiac catheterization in a supine position screened by an Philips UNIQ FD10 C-arm system platform. Eight milliliters of iodixanol-320 (contrast agent) was injected into each patient at every injection. The scanning parameters were: an X-ray tube voltage of 120–140 kVP, a field of view of 25 cm, and sequence lengths varying from 3 to 5 s at 14 frames per second. The resolution of each frame is 512×512.

2.2. Methods

A flow diagram of the proposed framework is shown in Fig. 2, which includes three major parts: contrast-filled frames selection based on U-Net, single frame stenosis detection based on an DSSD and false positive suppression based on seq-fps. The following three chapters will describe each part.

2.2.1. U-Net based contrast-filled frames selection

Automatic identification of the contrast-filled frames from an XCA sequence is the first step of the proposed algorithm. These frames are the most appropriate for stenosis detection since they show complete coronary artery structures. Inspired by previous works [17,18] that concentrated on detecting the contrast inflow, we employ a simple method based on U-Net, which is a classic neural network for image segmentation, to handle the contrast-filled frames selection task.

In an XCA sequence, a complete coronary artery structure gradually emerges as the radiopaque contrast agent flows in and subsequently fades as the contrast agent flows out. Since a well-trained U-Net only responds to visible contrast-filled vascular parts, we think that the area of segmented vascular structures in the output binary image can represent the overall contrast-filling degree of the input frame. Therefore, the most contrast-filled frame can be determined by searching for the maximum of the U-Net segmented vascular areas. However, a single frame is not enough since our algorithm requires the necessary temporal constraint from consecutive X-ray frames to reduce the number of false positives. Therefore, we select N frames before and after the most contrast-filled frame. According to clinical experience, with a reasonable value for N, all the $(2N + 1)$ frames are sufficiently contrast-filled and

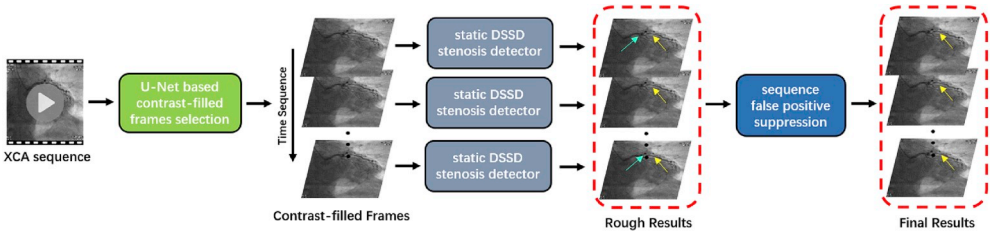


Fig. 2. Framework of the proposed method. The whole algorithm works as follows: first, the contrast-filled frames of an input XCA sequence are selected based on the U-Net segmentation results (shown in chronological order from top to bottom). Then, the DSSD provides rough results for each selected frame (yellow arrows for true positives and aqua arrows for false positives). Finally, the seq-fps module summarizes the rough results and removes false positives, generating the final results.

useful for stenosis detection and diagnosis. The DSSD network will generate rough stenosis detection results for each of these frames, preparing for the subsequent seq-fps module.

2.2.2. *Deconvolutional single-shot multibox detector*

The DSSD network (shown in Fig. 3) is the second part of the proposed framework. Apart from the powerful feature representation ability, the DSSD is also chosen for its high efficiency [19] and ability to be trained from scratch [20]. The backbone model of the applied DSSD is a typical VGG (Visual Geometry Group) [21] convolutional neural network, consisting of 19 convolutional layers with $3\times$ 3 kernels and 7 max pooling layers. Since the resolution of the input X-ray image is 512×512, the DSSD produces feature maps with resolutions ranging from 256×256 to 4×4. The deconvolution module (shown in Fig. 4) doubles the resolution of the high-level feature map with the learned deconvolutional layer and further combines feature maps from two different levels by elementwise summation. This step merges semantic information and location information, generating feature maps with richer contents that are beneficial for the detection task. Three deconvolution modules are applied to produce the final feature map with a resolution of 32×32. At the bottom of the network are two branches for classification and localization, which are also convolutional layers that adapt the channel number of the output feature map to the detection target. The localization branch outputs four channels with position information, and the classification branch outputs two channels with stenosis existence information for each default box. After the model is designed, we follow three basic procedures to train the DSSD.

First, we generate fixed bounding boxes called default boxes with different positions, scales and shapes. Taking an $n \times n$ feature map as an example, we regard it as $n \times n$ cells, whose centers are set as the default box centers. Then, the x and y coordinates of the default boxes' centers are: $(x,\ y) = \left(\frac{i+0.5}{n},\ \frac{j+0.5}{n}\right)$, $i, j = 0, 1, \ldots, n-1$, where i and j are incremental variables denoting the horizontal and vertical indexes,

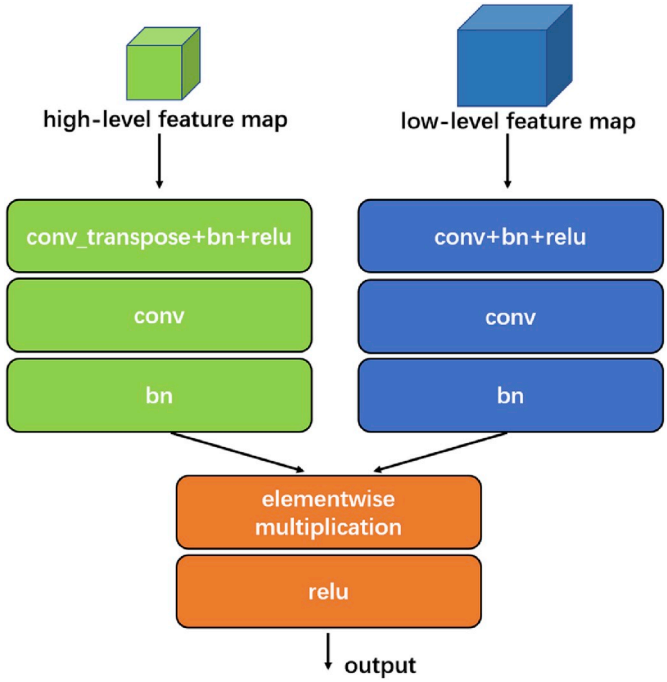


Fig. 4. Structure of the deconvolution block. The green block denotes a high-level feature map with lower resolution and the blue block denotes a low-level feature map with higher resolution. 'conv', 'bn' and 'relu' indicate convolution, batch normalization and the rectified linear unit, respectively; 'conv_transpose' denotes the transposed convolution that doubles the resolution of the high-level feature map.

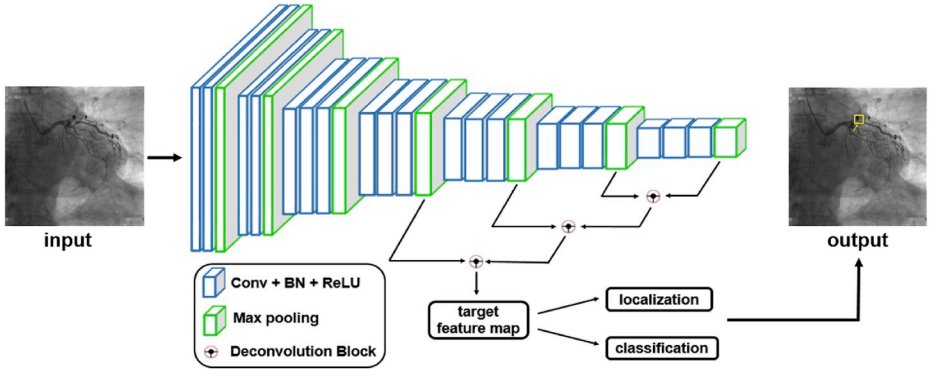


Fig. 3. Structure of the DSSD network. It requires a raw XCA frame as input and outputs the detected stenosis (marked in yellow rectangles). The blue block denotes the combination of three basic operations – 'Conv' (convolution), 'BN' (batch normalization) and 'ReLU' (rectified linear unit), and the green block denotes the max pooling operation. The localization and classification branches are two convolutional layers that determine the existence and position of a stenosis.

respectively, of a given cell. Furthermore, given a scale of s and an aspect ratio of r, the width (w) and height (h) of a default box are obtained: $(w, h) = \left(s\sqrt{r}, \frac{s}{\sqrt{r}}\right)$. A default box is thus denoted by (x, y, w, h).

Second, we select the positive default boxes and encode the localization information. For a default box, if its IoU (Intersection-over-Union) value with any annotated ground truth box exceeds the threshold $T_{\text{pos}-\text{iou}}$, it is regarded as positive. The IoU value is computed as follows:

$$IOU = D \cap G / D \cup G, \tag{1}$$

where D and G refer to the default box and ground truth box, respectively. Then, by computing the offsets from its best-matched ground truth box, a positive default box obtains the localization information from the ground truth labels. The four offsets of the x-coordinate, y-coordinate, width and height, corresponding to the four-channel output for each default box in the localization branch, are the regression targets of the DSSD. The offsets (denoted by O_*) are computed as follows:

$$O_x = c_1 (x_g - x_d) / w_d, \tag{2}$$

$$O_y = c_1 (y_g - y_d) / h_d, \tag{3}$$

$$O_w = c_2 log(w_g / w_d), \tag{4}$$

$$O_h = c_2 log(h_g / h_d), \tag{5}$$

where x, y, w and h refer to x-coordinate, y-coordinate, width and height, respectively. The subscript indexes d and g refer to the default box and ground truth box, respectively. c_1 and c_2 are two constants that adjust the scales of the offsets to accelerate network training.

Finally, we adjust the positions, scales and shapes of the positive default boxes towards their matched ground truths by iterative network training. The localization loss L_{loc} is defined as in the paper by Fu et al. [16]:

$$L_{\text{loc}}(l_i^m, g_i^m) = \sum_{i \in \text{p}} \sum_{m \in \{x, y, w, h\}} f_s(l_i^m - g_i^m), \tag{6}$$

$$s(z) = \begin{cases} 0.5z^2 & |z| < 1 \\ |z| - 0.5 & otherwise, \end{cases} \tag{7}$$

where f_s denotes the smooth-L1 function and p denotes the positive category. l_i^m and g_i^m denote the predicted and ground truth box offsets, respectively. In addition to the localization loss, the classification loss is also essential for network training. It should be noted that the localization loss is only for positive default boxes and the classification loss is for every box. However, considering that there are far more negative boxes than positive ones and that a sample imbalance is harmful to training, we randomly select negative samples to maintain a reasonable ratio. The classification loss L_{cls} is calculated with the binary cross-entropy:

$$L_{\text{cls}}(x_i, p_i) = \sum_i (x_i log p_i + (1 - x_i) \log(1 - p_i)), \tag{8}$$

where x_i and p_i denote the category label and predicted probability, respectively. The total loss L is the weighted sum of the classification loss L_{cls} and the localization loss L_{loc}:

$$L = \frac{1}{N_{\text{pos}}} (L_{\text{cls}} + \alpha L_{\text{loc}}), \tag{9}$$

where α is a weight parameter and N_{pos} denotes the number of positive default boxes.

When testing, the trained DSSD outputs a confidence score and offsets for each default box. We select the positive boxes and reverse the computational process of equations (1)–(4) to obtain real bounding boxes. Finally, NMS (non-maximum suppression) [22] is conducted to select the most suitable bounding box for a stenosis and remove redundant boxes.

2.2.3. Sequence-false positive suppression

The DSSD has high sensitivity; however, it is still influenced by a certain number of false positives. These false positives are generally stenosis-like structures generated by time-dependent contrast agent inhomogeneity and vessel motion, which might be unstable in the time domain. This phenomenon inspires us to exploit the potential temporal information of an XCA sequence to remove false positives.

Based on a video object detection algorithm seq-nms (sequence-non-maximum suppression) [23], we design the temporal module called seq-fps. It is performed on the DSSD network results of consecutive contrast-filled frames selected by U-Net, which selects the stenosis that most frequently appears in an XCA sequence, thus filtering out remaining random false positives. The module begins with the first two frames of the selected frames. A candidate box in the first frame can be linked to a candidate box in the second frame if their IoU value is above the threshold $T_{\text{seq}-\text{iou}}$. If two or more candidate boxes in the second frame satisfy the requirement, the box with the largest IoU value will be chosen to establish detection persistence. If no box linkage in the first two frames exists, the procedure will continue to the next two adjacent frames until a box linkage is finally found. Then, for the following frame, if the IoU value between one of its candidate boxes and the end of the linkage is greater than $T_{\text{seq}-\text{iou}}$, the box will join and lengthen the linkage, becoming the new end. A box linkage will be built frame by frame following these procedures. Multiple box linkages focusing on different detected stenoses can be established simultaneously. It is also possible that the detection of a specific stenosis will be interrupted at middle frames due to motion or overlaps, which causes the disconnection of an ideal box linkage. Therefore, we will search and reconnect these linkages. A more complete new linkage will be established based on two shorter linkages if the end box of one linkage and the start box of the other linkage satisfy the $T_{\text{seq}-\text{iou}}$ threshold requirement. Fig. 5 shows the process of the seq-fps module. In practical applications, N_{cf} denotes the number of selected contrast-filled frames and N_{sf} is the threshold for the established box linkage length. If a box linkage length is greater than N_{sf}, we regard its corresponding detections as true positives and preserve them. The remaining candidate boxes will be regarded as false positives and removed.

3. Experiments and results

3.1. Implementation details

We obtained 148 XCA sequences in total. Five-fold cross-validation experiments were conducted with 123 sequences: each time, 4/5 of the data are used for training, and 1/5 of the data are used for validation. The remaining 25 sequences are fixed as the test set. The validation set is used to tune the hyperparameters, and the test set is used to evaluate the performance of the final optimal model. The numbers of sequences and the corresponding numbers of stenoses are listed in Table 1.

When training the segmentation network U-Net, 60 images are collected from different sequences with experienced cardiologists offering pixel-level annotations to distinguish foreground vessel structures from the background. The gray values of the XCA images are normalized to -1-1 before being input into U-Net. When training the DSSD, each sequence is divided into single frames. The cardiologists randomly select 6 discontinuous contrast-filled frames from each sequence and provide ground truth bounding boxes for stenoses with the professional labeling software Colabeler. The selection of multiple frames can be regarded as a special data augmentation approach, which introduces the variance within the XCA sequence (e.g., motion and brightness variation) into the training set. According to clinical experience, only a stenosis with a degree≥50% is considered significant and requires treatment. When

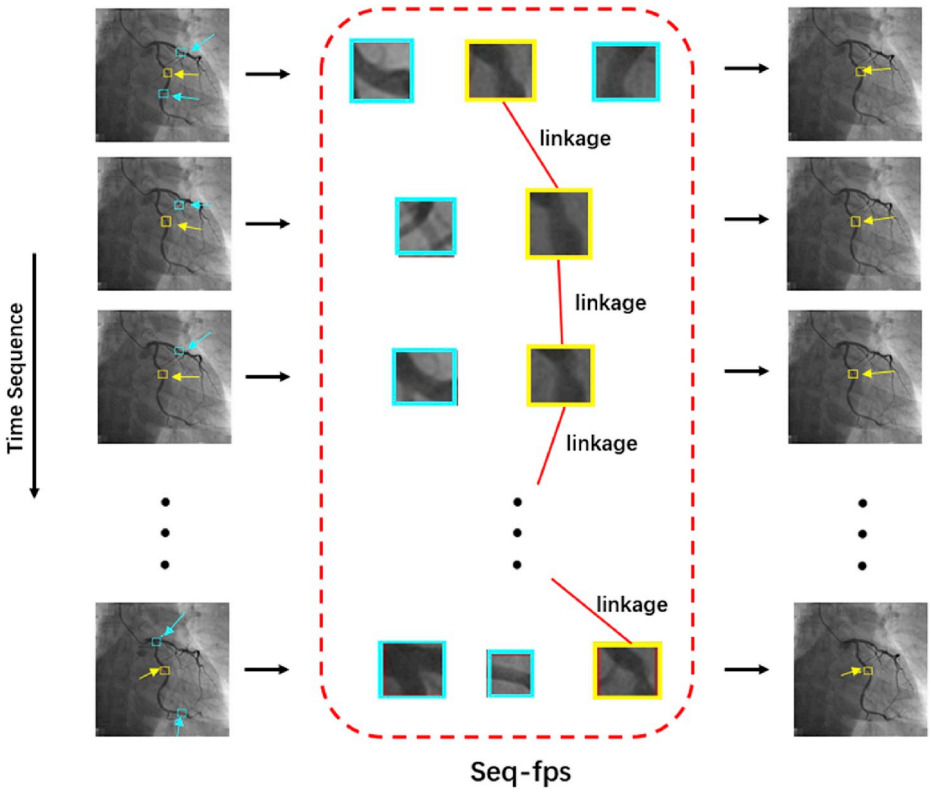


Fig. 5. Illustration of sequence false positive suppression. The images on the left are consecutive frames from an XCA sequence shown in chronological order from top to bottom, and the images on the right are their corresponding results after seq-fps (true positives and false positives are marked with yellow and aqua, respectively). The middle images are candidate detections provided by the DSSD for single frames, which are cropped from the images and zoomed for better visualization. Once the detected regions from two consecutive frames have an IoU value above the given threshold, a linkage is built for them (shown with a red line). The seq-fps module keeps all the candidate detections on the linkage and removes all the remaining detections.

Table 1
Numbers of sequences and the numbers of stenoses in the datasets ('Val' means validation).

Dataset	Val1	Val2	Val3	Val4	Val5	Test
Number of sequences	25	24	25	24	25	25
Number of stenoses	34	32	33	34	35	36

labeling, once a significant stenosis is identified, regardless of whether it is at the main vessel or at a bifurcation, a bounding box of suitable size is drawn to cover it. In our dataset, the stenosis size is usually approximately 30 × 30 pixels, taking up a small portion of the vessel segment. In cases where a long and thin stenosis appears, the provided ground truth bounding box only covers the junction of the stenosis and the normal vessel.

The DSSD network is implemented in the TensorFlow [24] framework. The software environment is: Tensorflow version 1.10.1, CUDA version 9.0.176, cuDNN version 7.0.3, and Python version 3.6.7, and the operating system is Ubuntu 16.04 LTS. The hardware environment is made up of a 4-core 2.4 GHz Intel Xeon E5-2630 v3 processor and a single Nvidia Titan X Pascal GPU. Convolutional layers are initialized with random weights sampled from a standard Gaussian distribution. The total number of iterations is 40K. The initial learning rate is 0.1, which decays exponentially by a power of 0.9. Batch normalization is conducted to achieve faster convergence and L2 regularization with $\lambda = 10^{-4}$ is employed to avoid overfitting. The batch size is 10, and the Adam optimizer is adopted to train the network. Data augmentation methods, which include horizontal and vertical flipping, rotation (90°, 180° and 270° clockwise), translation (30 or 60 pixels in the right, left, up and down directions), contrast adjustment ($\alpha = 0.5, 0.75, 1.5$) and brightness adjustment ($\beta = -50, +50$) with the formula $y = \alpha x + \beta$ are employed to expand the training set, making the network more robust to image variations in XCA sequences. The scale of the default box of the DSSD is set to 40 × 40 pixels and is normalized to 0.078125 (40/512), which is slightly larger than the real stenosis size to ensure full coverage. Three aspect ratios (1, 2/3 and 3/2) are arranged so there are three default boxes in one cell. We set $T_{\text{pos-iou}} = 0.5$, $c_1 = 5$, $c_2 = 10$, the loss weight $\alpha = 1$ and maintain a positive-negative ratio of 1:3 when training, as suggested in Ref. [25]. The threshold of the output confidence score is set to 0.9 to select positive predicted boxes when testing. The parameters of the seq-fps module are chosen with cross-validation experiments described in chapter 3.4.

In the following experiments, three metrics, the sensitivity (SN), positive predictive value (PPV) and F1-score, which are calculated with the number of TPs (true positives), FPs (false positives) and FNs (false negatives) are used to evaluate the performance of the algorithm:

$$SN = TP/(TP + FN), \tag{10}$$

$$PPV = TP/(TP + FP), \tag{11}$$

$$F1 - score = 2*SN*PPV/(SN + PPV). \tag{12}$$

3.2. Performance of U-Net for contrast-filled frames selection

Fig. 6(a) shows raw frames from an X-ray sequence shown in chronological order. Fig. 6(b) shows the corresponding U-Net segmentation results. As the contrast agent flows in, the coronary artery structure gradually becomes complete, and simultaneously, the corresponding segmented vascular area becomes larger. This figure qualitatively reveals U-Net's ability to select contrast-filled frames. To conduct a quantitative evaluation, we compare the U-Net-based method with four other methods. Among them, three are also segmentation-based methods, and one is a classification-based method.

The three segmentation-based methods are the Frangi filter algorithm [26] with Otsu's thresholding [27], Coye's method [28] and MSRG (multiscale region growing) [29]. Fig. 6(c), (d), and 6(e) show the corresponding segmentation results of these methods. Visually, U-Net outperforms the other segmentation methods with a superior ability to preserve tiny structures. To generate a normalized contrast-filling degree, the segmented vascular area of each frame is divided by the

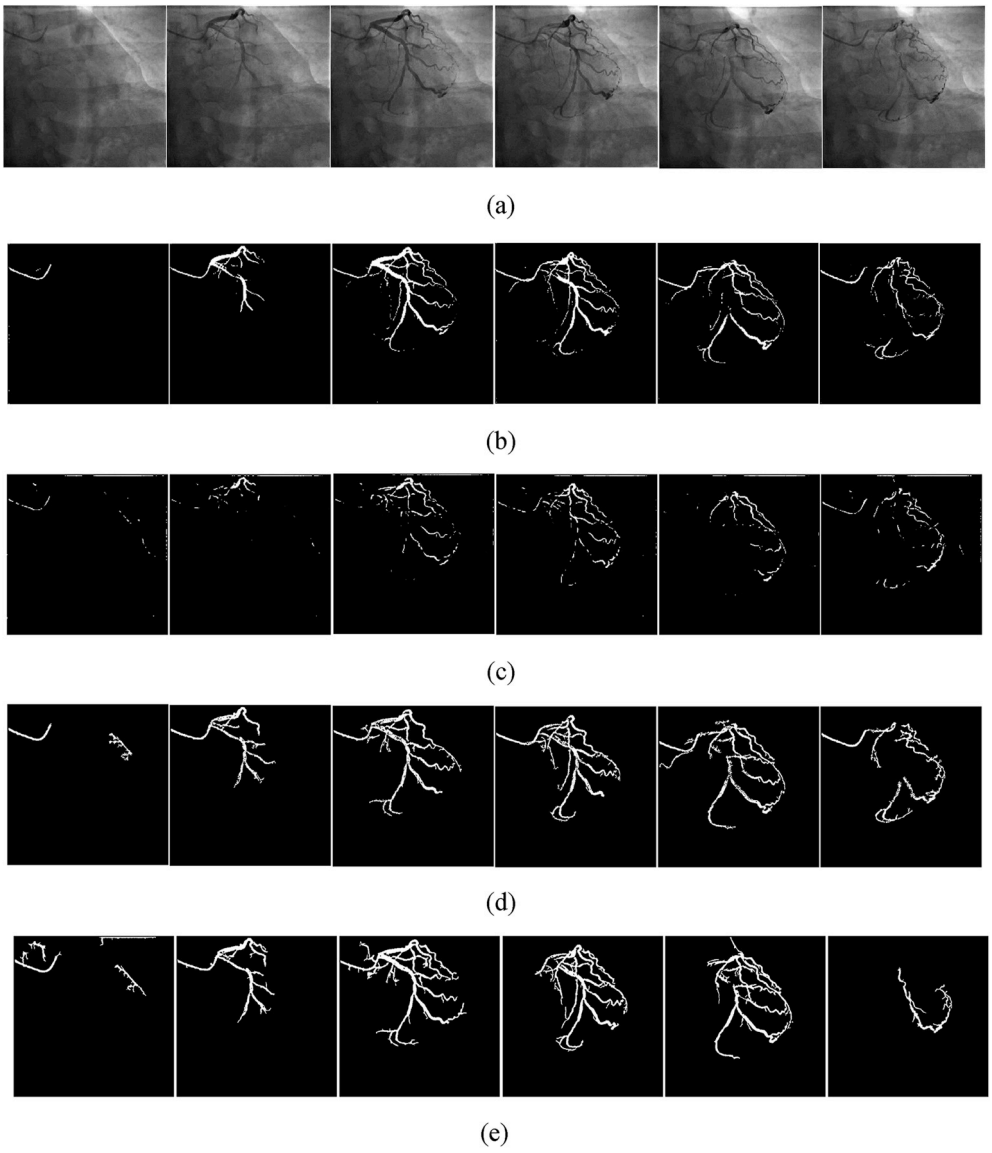


Fig. 6. Raw frames (a) from an X-ray sequence (64 frames in total) and their corresponding segmentation results: U-Net (b), Frangi + Otsu (c), Coye's method (d) and MSRG (e). These 6 frames are the 8th, 21st, 38th, 42nd,48th and 63rd frames, respectively, shown in chronological order.

maximum vascular area of the sequence.

In the classification-based method, we build a classification neural network ResNet [30] to deal with the frame selection task. It is trained with 90 XCA sequences with data augmentation. Instead of selecting one best frame from a sequence, which is difficult due to severe imbalance between the numbers of positive and negative samples, we label the frames as one of three categories ('zero', 'partial' and 'full') according to contrast-filling degree. The output probability of 'full' is used to represent the contrast-filling degree.

We used 45 XCA sequences to test these methods. Since a sequence usually has multiple contrast-filled frames that are equally effective for diagnosis, the cardiologists labeled three consecutive frames as the ground truth for each sequence. A selection within the ground truth is regarded as accurate. Fig. 7 shows the variations in the normalized contrast-filling degree with the frame index, based on the cases shown in Fig. 6. The U-Net curve is smooth and has an elegant trend, with a peak appearing at the 38th frame, corresponding to the ground truth. The other segmentation-based curves, although they have similar trends as that of U-Net, are much sharper, with peaks appearing at the 42nd, 48th and 36th frames. ResNet regards all the frames after the 32nd frame as contrast-filled frames, with the latter part of the curve being flat; therefore, the peak is the middle frame (i.e., the 48th frame; the middle frame of a flat curve is chosen as the peak). The quantitative comparison results, for the accuracy and speed under GPU acceleration are shown in Table 2.

3.3. Performance of the DSSD for single-frame detection

Fig. 8 shows the DSSD network results of single frames from three different patients. All the predicted stenoses are annotated automatically by the algorithm. For better visualization, we manually highlight the true positives with yellow rectangles and arrows and the false positives with aqua rectangles and arrows. We can see that the DSSD shows high sensitivity for stenosis detection but the existence of false positives remains a problem. As mentioned in chapter 3.2, for each sequence, a cardiologist selected three frames that are best for diagnosis. By averaging the DSSD network results of these selected frames, we obtain the single-frame detection performance of the validation sets with a sensitivity of 86.3%, a PPV of 46.5% and an F1-score of 60.4%.

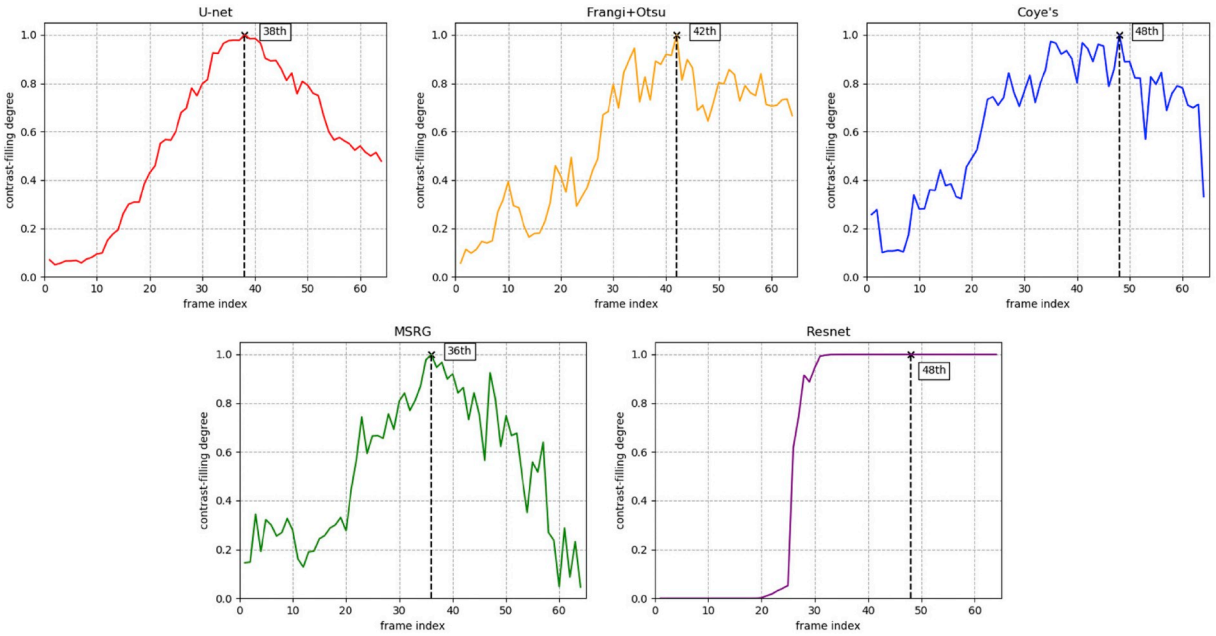


Fig. 7. Variations in the contrast-filling degree with the frame index (cases in Fig. 6). The horizontal axis is the frame index, and the vertical axis shows the normalized contrast-filling degree. The results of U-Net, Frangi + Otsu, Coye's method, MSRG and ResNet are shown.

Table 2
Quantitative comparison of the frame selection methods measured by accuracy and speed.

Method	Accuracy	Speed (seconds/frame)
U-Net	86.7%	0.035
Frangi + Otsu	44.4%	0.252
Coye's method	64.4%	0.207
MSRG	71.1%	0.463
ResNet	33.3%	0.017

3.4. Performance of the seq-fps module and parameter selection

There are three hyperparameters worth noting: the number of selected contrast-filled frames N_{cf}, the number of stenosis-appearing frames N_{sf} and the IoU threshold $T_{seq-iou}$, the latter of which determines whether bounding boxes from neighboring frames belong to the same stenosis-like region. The displacements of a specific stenosis between frames, mainly caused by cardiac motion, can be very large compared with its own tininess. Therefore, our experimental results suggest that $T_{seq-iou}$ should be relatively small to build stable stenosis linkages across target frames. It is set to 0.1 in our experiment, which works fairly well. We pay more attention to N_{cf} and N_{sf}. The average heart rate of an adult is approximately 75 beats per minute, so the cardiac cycle is approximately 0.8 s. Considering that the FPS (frames per second) of the collected XCA sequence is fixed to 14, approximately 11 (14*0.8 = 11.2) frames are needed to observe a complete cardiac cycle. However, variations exist among the heart rates in different XCA sequences; therefore, we try three different values of N_{cf} $(2N + 1)$: 9, 11, 13. Since N_{sf} should be smaller than N_{cf}, three N_{sf} values ($N_{cf}-1$, N_{cf}-2 and N_{cf}-3) are also arranged for each N_{cf}. We conduct cross-validation experiments to test the performance of different parameter combinations. Fig. 9 shows the average sensitivity, PPV and F1-score of the validation sets. We choose the optimal parameters by comparing the composite index F1-score. The variance analysis experiment is carried out with MATLAB (2014a) and generates $P = 4.70e-18 < 0.01$, showing that N_{cf} and N_{sf} have a significant influence on the detection performance. Fig. 9 demonstrates that the selection of $N_{cf} = 11$ and $N_{sf} = 8$ is the best choices, obtaining the highest F1-score of 84.2% with a sensitivity of 88.7% and a PPV of 80.2%. Further analysis of the influence of these two parameters will be presented in the discussion section.

Based on $N_{cf} = 11$ and $N_{sf} = 8$, we compare the detection results of the test set with and without seq-fps. Fig. 10 shows the detection results of patient #2 from Fig. 8. The single-frame detection results of the DSSD for 8 consecutive contrast-filled frames selected by U-Net are shown in Fig. 10 (a). A true stenosis is detected steadily, appearing at every frame, while false positives lack temporal persistence, appearing randomly. The corresponding seq-fps-modified results of the frames in Fig. 10 (a) are shown in Fig. 10 (b). False positives have been successfully removed while the true stenosis is still maintained. When testing with the whole

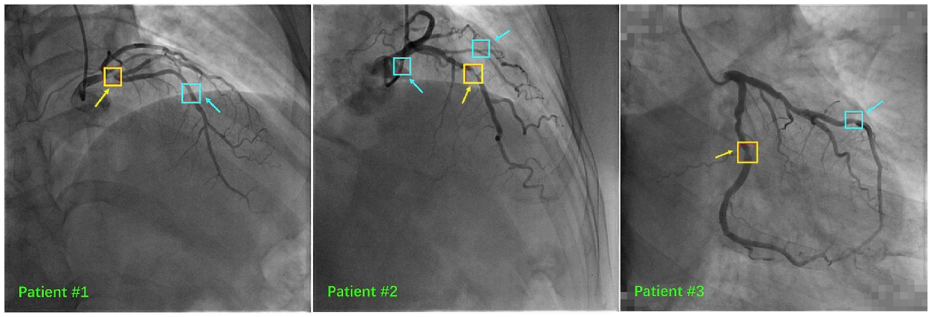


Fig. 8. Single-frame detection results of the DSSD from three patients. True positives and false positives are marked with yellow and aqua, respectively.

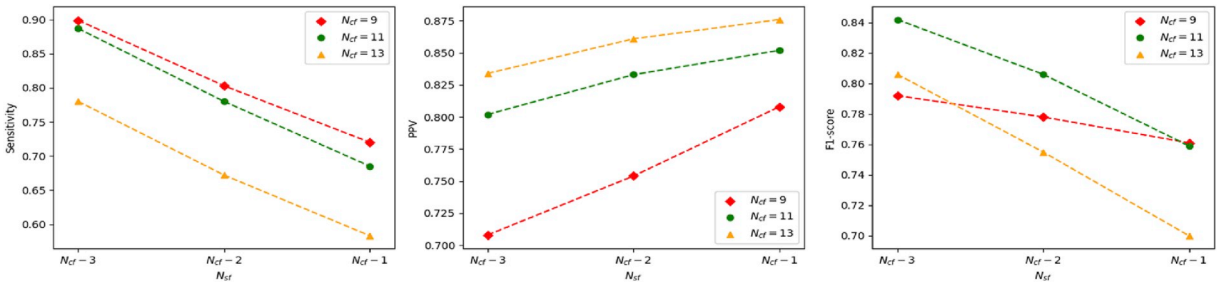
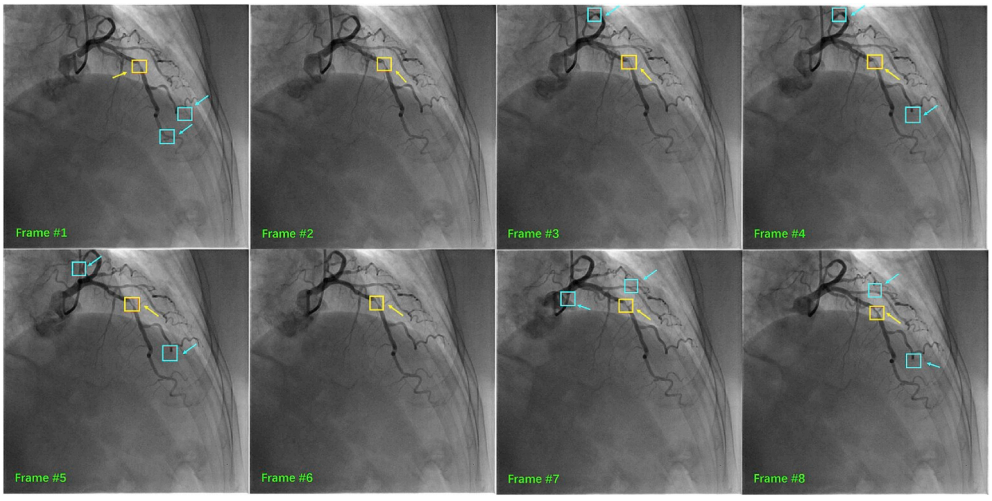
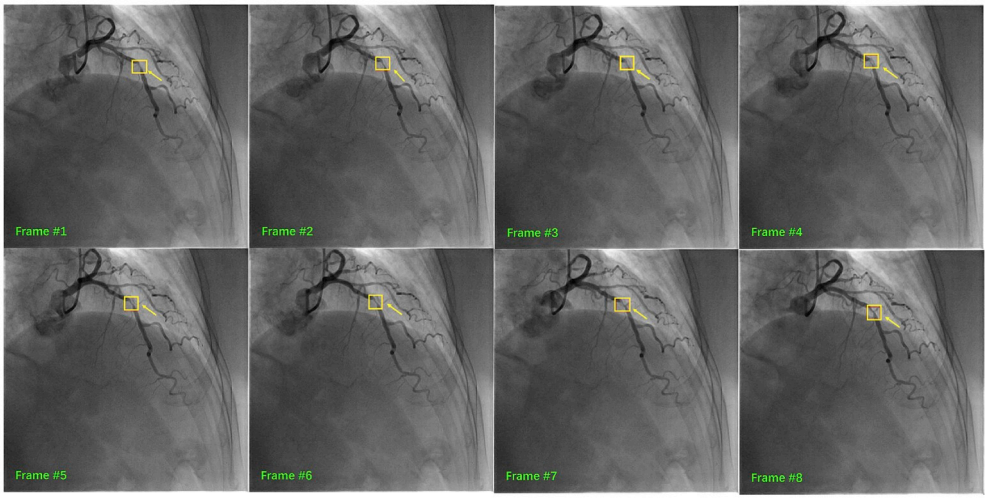


Fig. 9. Average sensitivity, PPV and F1-score of the validation sets with different N_{cf} and N_{sf} combinations.



(a)



(b)

Fig. 10. DSSD and seq-fps detection results for one XCA sequence. True positives and false positives are marked with yellow and aqua, respectively. (a) shows 8 consecutive frames from one sequence showing the DSSD-based single-frame detection results. (b) are the corresponding results of (a) after seq-fps processing with false positives removed.

dataset, we obtain average single-frame detection results with a sensitivity of 87.1%, a PPV of 46.1% and an F1-score of 60.3%. After the processing of the seq-fps module, the sensitivity remained almost unchanged, and the PPV greatly improved from 46.1% to 79.5%. This finding indicate the efficiency of the temporal constraint of the XCA sequence for reducing the number of false positives.

3.5. Robustness test

Due to various conditions of clinical data acquisition, the collected XCA sequences might have different illumination and contrast levels. Experiments have been conducted to verify the robustness of the proposed framework to these variations with the test dataset. Contrast and brightness variations are simulated with formula $g(x, y) = \alpha f(x, y) + \beta$. $f(x, y)$ and $g(x, y)$ denote the original and transformed gray values,

respectively, at pixel (x, y), and α and β adjust the contrast and brightness levels, respectively. We experimented with $\alpha = 0.5, 0.75, 1.25, 1.5$ and $\beta = -75, -25, +25, +75$. Fig. 11 shows the results of the robustness test for the case in Fig. 10 (we only show the detection result of frame #5 for simplicity). The overall results of the test dataset demonstrate that the original detections remain stable under most conditions. Even in the extreme situation $\beta = -75$ where images become very dark and vessel structures merge with the background, the framework still has good performance with only 3 true positives and 2 false positives missing compared with the original detection. Therefore, generally speaking, the proposed method is robust to image variations.

Apart from the robustness to image variations, the performance of the proposed method on healthy patients is also considered. We collected another 5 XCA sequences from patients without stenosis and tested the proposed method on them. The results demonstrate that the proposed framework does not generate any false detections.

3.6. Computational analysis

The proposed algorithm is implemented in the TensorFlow framework with Python. The computations are completed on a computer with a 4-core 2.4 GHz Intel Xeon E5-2630 v3 processor and a single Nvidia Titan X Pascal GPU. It takes U-Net 0.035 s to generate the segmentation results and the DSSD 0.040 s to generate the detection results for a single frame. The XCA sequences have an average length of 50 frames, so the algorithm requires 1.75 s for U-Net segmentation, 0.21 s for the selection of 11 contrast-filled frames, 0.44 s for single-frame detection and 0.49 s for sequence false positive suppression to process an XCA sequence. The total computation time is 2.89 s.

3.7. Comparison with existing methods

We compare our method with some of the existing methods for stenosis detection. Since related works based on XCA are limited and CTA is also a common tool for stenosis detection in clinical practice, the comparison includes CTA based methods. Considering that the detection result is more important than the imaging technique when diagnosing, we believe this comparison is of practical significance. The methods reported by Shahzad et al. [2] and Broersen et al. [4] are two outstanding methods from the 2012 MICCAI Challenge [31] that detect stenosis based on CTA while the one by Compas et al. [9] is based on XCA. Zreik et al. [5] recently proposed a new method based on recurrent neural network. Table 3 shows the comparison of these methods. The proposed method is superior in both the sensitivity and PPV. Further analysis can be found in chapter 4.3.

Table 3
Quantitative comparison of the stenosis detection methods measured by the sensitivity (SN), positive predictive value (PPV) and F1-score.

Method	SN	PPV	F1-score
Shahzad et al.	54.1%	26.8%	35.8%
Broersen et al.	27.7%	30.9%	29.2%
Compas et al.	86.4%	57.6%	69.1%
Zreik et al.	80.0%	70.6%	75.0%
Proposed method	**87.2%**	**79.5%**	**83.2%**

4. Discussion

4.1. Analysis of the weak temporal persistence of false positives

Fig. 10 (a) shows the DSSD network results of 8 single, consecutive frames from one XCA sequence. It can be observed that the true stenosis is consistently detected, while false positives might be quite different even in neighboring frames. In other words, false positives have weak temporal persistence. We analyze this phenomenon and offer explanations. True stenoses usually have relatively strong feature representations on X-ray angiograms, such as the unexpected narrowing of the vessels and grayscale value variations compared with neighboring vascular areas. Therefore, they are relatively easier to be detected regardless of the complex variations such as vascular structure motion or contrast agent flow in an XCA sequence, showing good temporal persistence. In contrast, false positives have weaker features. They exist because the detector is interfered with by factors such as curved vessels and ribs in the background. However, these interference factors vary across frames, generating different false positives. For example, a false positive caused by instantaneous contrast agent inhomogeneity might appear only in one frame, while a false positive caused by vessel motion only appears in another frame. Therefore, in the DSSD network results, false positives seem to appear irregularly. In conclusion, false positives are sensitive to variations in the XCA sequence, which makes them have weak temporal persistence. This phenomenon also inspires us to exploit temporal information and suppress false positives.

4.2. Analysis of parameter selection

The proposed framework has many parameters, and we divide them into two categories. The first category refers to the traditional parameters for building the detection network, for example, the aspect ratios, $T_{\text{pos-iou}}$, and loss function weight. Among them, we specifically discuss the influence of $T_{\text{pos-iou}}$ and its important function. $T_{\text{pos-iou}}$ is an IoU threshold determining whether a default box is positive or not. If we

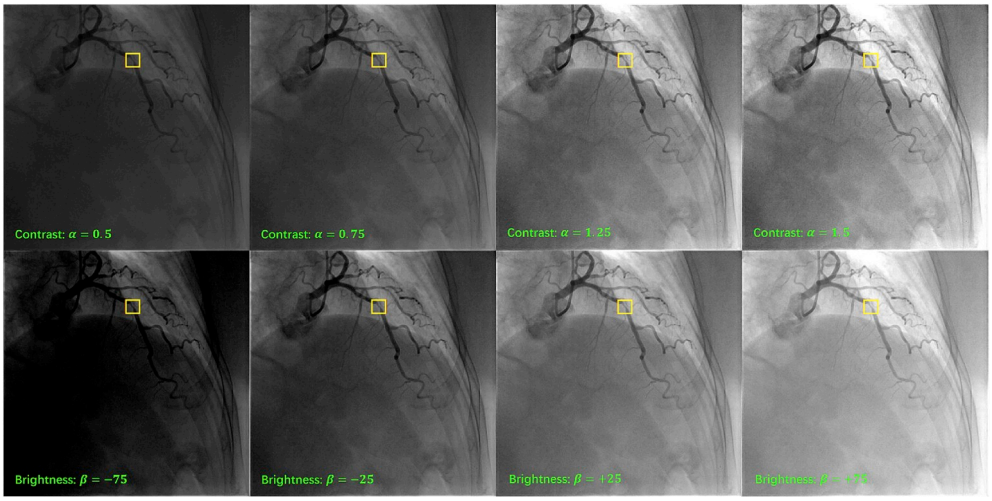


Fig. 11. Robustness test results under different contrast and brightness levels (frame #5 in Fig. 10).

increase $T_{pos-iou}$, fewer positive bounding boxes will be generated. Therefore, the average distance between positive boxes and their corresponding ground truth will be smaller, which will make position regression easier. However, it might also lead to network overfitting since the number of regression targets decreases. Conversely, if we decrease $T_{pos-iou}$, the number of positive boxes will increase. More default boxes that are far away from the ground truth will be considered positive, which is harmful to the classification performance of the network. We set $T_{pos-iou}$ to 0.5 in our experiment, which has been validated as effective in many existing object detection frameworks. Further research into the adjustment of $T_{pos-iou}$, together with other traditional parameters, will be conducted as future work of this study.

The second category refers to the unique parameters in the designed module. Among them, we think N_{cf} and N_{sf} of the seq-fps module are the most important. N_{cf} decides how many contrast-filled frames will be selected for temporal processing and N_{sf} decides whether a box linkage generated by the seq-fps module should be preserved. If we decrease N_{cf} or N_{sf}, we weaken the temporal constraint from the X-ray sequence and the temporal detection approaches single-frame detection. Therefore, the number of missed stenoses will be reduced, but the number of false detections will increase. Conversely, if we increase N_{cf} or N_{sf}, false negatives will become the major problem. Larger values will involve more uncertainties for temporal processing and require more accurate former DSSD network results. Theoretically, N_{cf} can cover all contrast-filled frames and N_{sf} can be equal to N_{cf}. However, such a selection might not be reasonable due to its extremity. In chapter 3.4, we have offered a theoretical basis for reasonable N_{cf} and N_{sf} selection: a whole cardiac cycle must be observed. A further cross-validation experiment is conducted to determine their specific values. Variations in the experimental results with different parameter combinations in Fig. 9 have confirmed our analysis above and demonstrate that $N_{cf} = 11$ and $N_{sf} = 8$ is the best parameter combination that achieves good sensitivity while suppressing false positives.

4.3. Analysis of the comparison experiment

Among the comparison methods, the methods reported by Shahzad et al. [2], Broersen et al. [4] and Compas et al. [9] are similar; all of which detect stenosis by measuring vessel diameters. Vascular structures are enhanced and segmented with various traditional algorithms, and then the vessel diameters are measured along the centerline. A stenosis is identified where the diameter has an abnormally small value. Therefore, the final detection result is highly dependent on the preprocessing results. Inaccurate segmentation results easily lead to false positives and the three methods have relatively low PPVs: 26.8% [2], 30.9% [3], and 57.6% [4], respectively. The method proposed by Zreik et al. [5], a relatively new method, extracts patches along the centerline and uses a recurrent neural network to generate features for identifying stenosis. It obtains good results while achieving a balance between the sensitivity (80.0%) and PPV (70.6%). However, with this method, the original curved coronary arteries are straightened with medical software. This structural variation might produce potential errors. In the proposed method, we deal with stenosis detection in a different way by treating it as an object detection task. For a single X-ray angiogram, the detection is directly conducted on the raw image by an object detection network without preprocessing procedures in other comparison methods. Moreover, we have tried to exploit the temporal information of the XCA sequence and designed the seq-fps module, which greatly suppresses false positives. Finally, the proposed method outperformed other comparison methods, achieving relatively higher sensitivity of 87.2% and PPV of 79.5%.

5. Conclusions

Automatic detection of coronary artery stenosis in X-ray angiograms is a significant but challenging task. In this study, we present a deep learning-based method with temporal constraints on the X-ray sequences to detect stenosis. U-Net selects contrast-filled frames that are most beneficial for subsequent stenosis detection. The DSSD then offers rough detection results for these single frames, which reveals high sensitivity. However, false positives still exist. The seq-fps module exploits the temporal information of the XCA sequence to suppress false positives and generate the final results. The proposed method achieves sensitivity of 87.2% and PPV of 79.5%, outperforming existing methods. The experimental results demonstrate that the proposed method has the potential to establish a computer-aided diagnosis system to automatically detect stenosis, assisting clinical examinations.

Considering the limitations of this study, future works are threefold. First, we will establish a larger dataset by cooperating with more hospitals to collect clinical data with more variations. Second, we will further improve the robustness of the proposed method by performing research into parameter selection and even developing a self-adaptive method, which will help the proposed framework address the complex variations in the clinical environment. Finally, we will pay attention to not only the existence but also the property and degree of a stenosis. These findings will lead to SYNTAX score [32], which is an important index in clinical diagnosis.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Declaration of competing interest

The authors declares that they have no conflict of interest.

Acknowledgements

This research is partially supported by the National Key research and development program (2016YFC0106200), Beijing Municipal Natural Science Foundation (L192006) and the funding from IMR of SJTU as well as the 863 national research fund (2015AA043203).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2020.103657.

References

[1] I. Abubakar, T. Tillmann, A. Banerjee, Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013, Lancet 385 (9963) (2015) 117–171.
[2] R. Shahzad, H. Kirisli, C. Metz, H. Tang, M. Schaap, L. van Vliet, W. Niessen, T. van Walsum, Automatic segmentation, detection and quantification of coronary artery stenoses on CTA, Int. J. Cardiovasc. Imag. 29 (8) (2013) 1847–1859.
[3] C. Wang, R. Moreno, O. Smedby, Vessel segmentation using implicit model-guided level sets, in: MICCAI Workshop" 3D Cardiovascular Imaging: a MICCAI Segmentation Challenge", Nice France, 1st of October 2012, 2012.
[4] A. Broersen, P. Kitslaar, M. Frenay, J. Dijkstra, Frenchcoast: fast, robust extraction for the nice challenge on coronary artery segmentation of the tree, in: Proc.of MICCAI Workshop" 3D Cardiovascular Imaging: a MICCAI Segmentation Challenge, 2012.
[5] M. Zreik, J. Wolterink, T. Leiner, M. Viergever, I. Isgum, et al., A recurrent cnn for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography, in: IEEE Transactions on Medical Imaging, 2018.
[6] J. Brieva, M. Galvez, C. Toumoulin, Coronary extraction and stenosis quantification in x-ray angiographic imaging, in: The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 1, IEEE, 2004, pp. 1714–1717.
[7] M.R. Fatemi, S. Mirhassani, E. Ghasemi, Detection of narrowed coronary arteries in x-ray angiographic images using contour processing of segmented heart vessels based on hessian vesselness filter and wavelet based image fusion, Int. J. Comput. Appl. 36 (9) (2011) 27–33.

[8] T. Wan, H. Feng, C. Tong, D. Li, Z. Qin, Automated identification and grading of coronary artery stenoses with x-ray angiography, Comput. Methods Progr. Biomed. 167 (2018) 13–22.
[9] C.B. Compas, T. Syeda-Mahmood, P. McNeillie, D. Beymer, Automatic detection of coronary stenosis in x-ray angiography through spatio-temporal tracking, in: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), IEEE, 2014, pp. 1299–1302.
[10] J. Dai, Y. Li, K. He, J. Sun, R-fcn: object detection via region-based fully convolutional networks, Adv. Neural Inf. Process. Syst. (2016) 379–387.
[11] R. Girshick, Fast r-cnn, Proc. IEEE Int. Conf. Comput. Vis. (2015) 1440–1448.
[12] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, Proc. IEEE Int. Conf. Comput. Vis. (2017) 2961–2969.
[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, Proc. IEEE Conf. Comput. Vis. Pattern Recog. (2016) 779–788.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. (2015) 91–99.
[15] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[16] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, Dssd: Deconvolutional Single Shot Detector, 2017 arXiv preprint arXiv:1701.06659.
[17] T. Chen, G. Funka-Lea, D. Comaniciu, Robust and Fast Contrast Inflow Detection for 2d X-Ray Fluoroscopy, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2011, pp. 243–250.
[18] H. Ma, P. Ambrosini, T. van Walsum, Fast prospective detection of contrast inflow in x-ray angiograms with convolutional neural network and recurrent neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 453–461.
[19] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., Speed/accuracy trade-offs for modern convolutional object detectors, Proc. IEEE Conf. Comput. Vis. Pattern Recog. (2017) 7310–7311.
[20] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, X. Xue, Dsod: learning deeply supervised object detectors from scratch, Proc. IEEE Int. Conf. Comput. Vis. (2017) 1919–1927.
[21] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:1409.1556.
[22] A. Neubeck, L. Van Gool, Efficient non-maximum suppression, in: 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, IEEE, 2006, pp. 850–855.
[23] W. Han, P. Khorrami, T.L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, T.S. Huang, Seq-nms for Video Object Detection, 2016 arXiv preprint arXiv: 1602.08465.
[24] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A System for Large-Scale Machine Learning, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}16), 2016, pp. 265–283.
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single Shot Multi-Box Detector, European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[26] A.F. Frangi, W.J. Niessen, K.L. Vincken, M.A. Viergever, Multiscale Vessel Enhancement Filtering, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 1998, pp. 130–137.
[27] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[28] Coye Tyler, Novel retinal vessel segmentation algorithm: fundus images, MATLAB Central File Exchange, https://www.mathworks.com/matlabcentral/fileexchange/50839-novel-retinal-vessel-segmentation.algorithm-fundus-images, 2019.
[29] A. Kerkeni, A. Benabdallah, A. Manzanera, M.H. Bedoui, A coronary artery segmentation method based on multiscale analysis and region growing, Comput. Med. Imag. Graph. 48 (2016) 49–61.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[31] H. Kirisli, M. Schaap, C. Metz, A. Dharampal, W. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M.A. de Graaf, M. Meijs, et al., Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography, Med. Image Anal. 17 (8) (2013) 859–876.
[32] G. Sianos, M.-A. Morel, A.P. Kappetein, M.-C. Morice, A. Colombo, K. Dawkins, M. van den Brand, N. Van Dyck, M.E. Russell, F.W. Mohr, et al., The syntax score: an angiographic tool grading the complexity of coronary artery disease, EuroIntervention 1 (2) (2005) 219–227.