Towards accurate coronary artery calcium segmentation with multi-scale attention mechanism

Yang Ning1 | Yunfeng Zhang2,3 | Xuemei Li1 | Caiming Zhang1,3,4

1 School of Software, Shandong University, Jinan, China
2 School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China
3 Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan, China
4 Shandong Co-Innovation Center of Future Intelligent Computing, Yantai, China

Correspondence
Caiming Zhang, School of Software, Shandong University, Jinan, China.
Email: czhang@sdu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61572292, 61332015, 61602277

Abstract
Coronary artery calcium is a strong and independent marker of atherosclerosis and cardiovascular disease. Typically, the accurate segmentation of computed tomography images of the chest is an important prerequisite and basis for coronary artery calcium identification and analysis. However, this is very challenging in practice because the boundaries of coronary artery calcium, the small lesions with large shape variation, are very blurry, resulting in poor performance in existing studies. To tackle this challenge, we present a novel Attention-based Multi-Scale Network called AMSN, which can process information through both the main and boundary branches in parallel. Key to our AMSN is a new non-local multi-scale context encoder module, which is mainly composed of the multi-scale attention mechanism and local global long short-term memory module. By aggregating the multi-scale context information, i.e. high-resolution low-level and low-resolution high-level features, the model's feature representative capability and deployment ability are improved effectively. Besides, we introduce a new boundary preserving loss, which can consider the boundary information of all coronary artery calcium together and establish links for the segmentation of different coronary artery calcium simultaneously. Extensive experiments demonstrate our AMSN enables reliable accurate coronary artery calcium segmentation for assisted cardiovascular disease diagnosis clinically.

1 | INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death in Western countries [1], with a higher risk than tumours and other diseases, and accounts for more than 40% of deaths from the disease in the population, ranking among the top 10 causes of death worldwide. Studies have shown that while some progress has been made in the prevention of CVD, serious challenges remain. Typically, coronary artery calcium (CAC) and coronary artery stenosis are the main causes of CVD. However, since the former is an important marker of atherosclerosis, it is usually the preferred option for detecting CVD. More specifically, since coronary calcium scoring can objectively reflect the degree of coronary atherosclerosis, the measurement of it is generally used in clinical practice to achieve rapid detection of CVD [2]. However, the most important step in the calculation of coronary calcium scoring is the segmentation of CAC.

In clinical practice, manual segmentation by a radiologist is currently the standard method of organ segmentation. However, manual segmentation is often time-consuming and laborious, and its results may vary from person to person, making it often impractical in the face of the explosion of computed tomography (CT) images. Automated segmentation based on CT images of the chest can effectively reduce the heavy workload of radiologists, and is also of great importance for the realization of auxiliary analysis and diagnosis of CVD. However, there are many practical difficulties to be faced in the automatic segmentation of CAC [3], firstly, CAC is usually small and morphologically variable; second, CT images usually have a complex background, which severely affects the accurate segmentation of CAC; besides, the degree of calcification distribution varies greatly from patient to patient; and finally, different scanning devices result in significant variability between different CT images. These challenges have resulted...
in some of the conventional methods being ineffective at the accurate segmentation of CAC. Therefore, a fast and accurate automated segmentation technique for calcification in chest CT images is essential to facilitate the diagnosis of CVD.

Over the past decades, many CAC segmentation methods based on chest CT images have been proposed, and they have addressed, to some extent, the inadequacy of manual segmentation with good results. Wu et al. and Ding et al. [4, 5] first segmented or roughly located larger tissues (e.g. the heart) to obtain a region of interest (ROI), which is used to obtain an automatic score for CAC. Refs. [6, 7] first rendered the coronary arteries by segmentation and then searched for high-intensity gradient transformations to detect CAC as a means of obtaining an automatic score for CAC. Eilot and Goldenberg [6] enable the detection, segmentation and scoring of CAC by modelling the image intensity distribution of coronary arteries, which reduces the risk of patient exposure to radiation and the cost of healthcare resources. Recently, deep learning-based methods have shown tremendous potential to assist in the diagnosis of CVD. Wolterink et al. [8] proposed a two-stage algorithm that first uses a convolutional neural network (CNN) to classify CAC regions; then reclassifies them based on RF as a way to suppress false-positive samples. Wolterink et al. [9] improved the above algorithm by, first, obtaining the ROI regions based on target detection and classifying only the pixels within the region, and then reclassifying the CAC regions based on the improved CNN. Although these methods improve the detection accuracy of CVD to a certain extent, they only assist in diagnosis by scoring CAC and do not achieve pixel-level segmentation of them. Therefore, improvements to the above methods are still needed to further facilitate their application to clinical settings.

Nowadays, the attention mechanism has shown powerful performance in extracting features with strong representative capabilities, which is applied to various tasks, such as target detection and image segmentation. OCNet [10] used the attention mechanism for the first time in the field of semantic segmentation, and with good results. It introduces the concept of object semantics and uses object context pooling strategies to generate more robust contextual information for semantic segmentation. DANet [11] can be seen as an improvement on OCNet [10], which first acts the attention mechanism on both the spatial and channel dimensions by integrating local dependencies with full global dependencies. A more robust feature expression is then obtained by aggregating the outputs of the two sets of attentional modules, which helps to obtain more accurate segmentation results. CCNET [12] proposes a new type of cross-attention module that effectively reduces the amount of computational complexity and memory footprint of the algorithm, as opposed to the approach that requires updating each pixel’s contextual information using all pixels, which is done using only the pixels on the cross-path.

This paper proposes a novel attention-based multi-scale network (AMSN) for accurate CAC segmentation on chest CT images, which integrates the main branch and the boundary branch designed to extract boundary constraint information into an end-to-end learning system. Our AMSN mainly consists of the following three aspects. First, we construct an autoencoder based on dilated convolution to reduce the number of network parameters, which can not only maintain the spatial resolution of the CT image but also obtain a larger receptive field; second, we proposed a new non-local multi-scale context encoder (NMC) module, which seamlessly connects the newly designed multi-scale attention mechanism with the local global long short-term memory (LG-LSTM) module, so that the pixel-level segmentation network can apply local guidance from the neighborhood location and global guidance from the full map to each pixel, which can effectively mine the complex short- and long-range spatial dependencies. Finally, we propose a hybrid loss consisting of a Laplace smoothing-based dice similarity coefficient (DSC) loss and a boundary preserving loss, which further improves the boundary segmentation accuracy of CAC. Extensive experiments show that our AMSN can effectively extract both the local and the global dependencies and boost the precise boundary segmentation of CAC.

In summary, our main contributions are listed below.

- For the first time we present a non-local multi-scale context encoder, by combining high-resolution low-level features and low-resolution high-level features; the model can learn the multi-scale scale contextual information, which effectively improves the boundary segmentation accuracy of CAC.
- By integrating the multi-scale spatial attention mechanism and the LG-LSTM module into a whole, the model can simultaneously model the short and long dependencies, which effectively enhances the feature representative capability.
- A new boundary preserving loss that can consider all CAC boundary information simultaneously is proposed. By combining it with the Laplace smoothing-based DSC loss, the model’s generalization ability and training stability are effectively improved.
- We construct a new dataset to assist in the diagnosis of CVD. Extensive experiments show that our AMSN is effective on this dataset and outperforms existing state-of-the-art methods by a large margin.

## 2 | METHODOLOGY

### 2.1 | Candidate region generation

To reduce the computational complexity of the CNN network and the interference of irrelevant background regions with CAC segmentation, all chest CT images will be preliminarily processed. For CAC segmentation, we usually have access to a range of prior knowledge. These include the fact that the identified lesions are coronary artery structures, the volume of the CAC region is usually not less than 1.5 mm$^3$ and the intensity value is larger than a standard threshold of 130 Hounsfield Unit (HU). Based on these analyses, we can improve the segmentation performance by roughly localizing the CAC region as follows.

Figure 1 illustrates the pipeline of candidate region generation, which mainly lies in the following six steps: 1) Take the original CT image (256 × 256) as input; 2) Based on the prior knowledge that the identified lesions must belong to the...
coronary artery, we used a standard threshold above 130 HU and morphological operations to roughly localize the CAC region to obtain the over-segmented CT image and the labeled image (256 × 256).

3) Use Equation (1) to get the coordinates of the target label based on the original encoded coordinates, which is encoded following the definition of region proposal networks (RPNs) [13].

\[
\begin{align*}
    t_x &= \frac{(x - x_a)}{w_a}, & t_y &= \frac{(y - y_a)}{h_a} \\
    t_w &= \log\left(\frac{w}{w_a}\right), & t_h &= \log\left(\frac{h}{h_a}\right)
\end{align*}
\]

(1)

where \(x, y, w, h\) denote the centre coordinates of the bounding box, width and height, respectively. \(x_a, y_a, w_a, h_a\) denote the center coordinates of the anchor, width and height, respectively.

4) Train the RPN to get the proposal area, and apply non-maximum suppression to obtain the barycentric coordinate (red star) of the final proposal area; 5) Map the barycentric coordinate back to the original CT image; and 6) To provide enough contextual information and reduce the computational complexity of the subsequent segmentation network, we generate candidate CT image and its corresponding labelled image (128 × 128) based on the barycentric coordinate by cropping the original CT image (red dotted square) and the labelled CT image. Note that, if the cropping area exceeds the boundary of the original CT image, its mirror image is used as a supplement. Experiments show that the generated candidate regions can cover CAC with nearly 100% recall among almost all used cases.

2.2 The framework of AMSN

As shown in Figure 2, we propose an AMSN that includes both the main branch and boundary branch. First, to obtain enough contextual information and reduce model parameters to facilitate model deployment, the upstream branch of the network, we refer to it as main branch, is composed of an auto-encoder based on the dilated convolutions; second, inspired by the spatial attention mechanism and the idea of low-order reconstruction, the downstream branch of the network is mainly composed of the NMC modules, which effectively enhance the representative capability of the lower level feature by using the higher level semantic feature as the boundary constraints of the newly designed multi-scale attention mechanism.

Since CAC usually have a small spatial resolution and blurred boundaries, sufficient contextual information is crucial for their segmentation. Generally, for the segmentation network to obtain enough context information, the receptive field corresponding to its convolution kernel should be large enough. However, for conventional CNN, the larger receptive field always means the less spatial resolution information, and such paradox is very unfavourable for the segmentation of CAC with a small spatial resolution. To allow for a larger receptive field while maintaining the spatial resolution, we employ the dilated convolution [14–16] with the dilation rate \(r\), which is equivalent to adding \(r - 1\) zeros between successive values of the regular convolution kernel. Besides, dilated convolution can also effectively reduce the parameters of the model. By continuously stacking dilated convolution layers, the receptive field grows exponentially while the parameter of the model only increases linearly [17].

Although the dilated convolution-based main branch can reduce the number of model parameters while maintaining the performance of the model, to further improve the precise boundary segmentation of CAC and the representative capability of the extracted feature, we have also introduced another branch, we refer to it as boundary branch, which further enhances the boundary segmentation accuracy of CAC by fusing the multi-scale attention mechanism with the idea of low-rank reconstruction.
FIGURE 2 Illustration of the architectural details of our attention-based multi-scale network. It consists of a main branch and a boundary branch, which mainly includes the non-local multi-scale context encoder module.

FIGURE 3 Illustration of the non-local multi-scale context encoder.

2.2.1 NMC module of the boundary branch

The spatial attention mechanism derives from human visual attention, and it is a mechanism for processing visual information formed during human evolution. Such a mechanism assumes that people’s attentional resources are finite, so the brain automatically distills the most useful information to guide human behaviour. For computer vision, the spatial attention mechanism has demonstrated powerful feature extraction capabilities in many tasks, such as semantic segmentation [18, 19], group activity recognition [20], salient object detection [21] and image restoration [22]. Based on the above analysis, we propose a new NMC module based on the spatial attention mechanism to further improve the feature extraction ability and boundary segmentation accuracy of our AMSN.

The traditional spatial attention mechanism [23] directly applies large matrix multiplication to capture long-range dependencies, which inevitably introduces excessive computational complexity. Inspired by the idea of low-rank reconstruction, the computational complexity is effectively reduced by fusing the low-level high-resolution semantic features with the high-level low-resolution semantic features using lightweight computation and memory while enhancing local features with semantic boundary constraints.

As shown in Figure 3, given local features \( X \in \mathbb{R}^{H \times W \times C} \) and \( Y \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C'} \) from the low-level and the high-level convolution layers as inputs to the NMC module, respectively. First, we apply two convolution layers with \( 1 \times 1 \) filters on \( X \) and \( Y \) to generate two feature tensors \( Q \in \mathbb{R}^{H \times W \times C'} \) and \( K \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C'} \), respectively, where \( C' \) is the channel number of feature tensors. Second, we reshape \( Q \) to \( \mathbb{R}^{HW \times C'} \) and \( K \) to \( \mathbb{R}^{\frac{HW}{4} \times C'} \). Finally, we employ a matrix multiplication between \( Q \) and the transposed \( K \), and perform a softmax operation to generate the low-rank attention map \( A \in \mathbb{R}^{\frac{HW}{4} \times \frac{HW}{4}} \) as follows:

\[
a_{ji} = \frac{\exp(Q \cdot K)}{\sum_{i=1}^{\frac{HW}{4}} \exp(Q \cdot K)_{ii}} \tag{2}
\]

where \( j \) represents the index of query positions, and \( i \) lists all possible positions. \( a_{ji} \) define the relationship between \( i \)th and...
\[
LV_j = \begin{cases}
Wv_j \cdot V + b, \quad &\text{if } j = \min(\mathbb{N}) \\
\sum_{i=1}^{j} (a_h \cdot V_i) + X_j, \quad &\text{otherwise}
\end{cases}
\]

where \( Wv \) represents the trainable transformation matrix, \( V \) and \( b \) are the trainable transformation matrix, \( Wv \), and \( b \) are the trainable transformation matrix, \( \ell \) is the channel number. The Hilbert curve-based serialization helps improve the segmentation performance of our AMSN.

In summary, the main difference between our multi-scale NMC module and traditional spatial attention mechanism is as follows: 1) Using both the low-level and the high-level semantic features as inputs to the spatial attention mechanism; 2) Reducing the GPU memory usage of the NMC module based on the idea of low-rank reconstruction; and 3) By integrating the multi-scale attention mechanism and the LG-LSTM module into a unified module, which can seamlessly model the relationship between long-range and short-range dependencies.

2.2.2 LG-LSTM module of the boundary branch

To endow the NMC module the ability to model the long-range and short-range dependencies simultaneously, we seamlessly fused the LG-LSTM module with the newly designed multi-scale attention mechanism. Figure 4 shows the overall structure of the LG-LSTM module, which consists of a series of LSTM units. Since the sequence of the patch-wise CT image patches is critical for LSTM performance, we use the Hilbert curve-based serialized feature sequence as the input to the LG-LSTM module. Unlike directly serializing the feature map horizontally or vertically, Hilbert curve-based serialization can better preserve the spatial locality of the CT image patches [24]. Intuitively, if we perform the serialization operation directly based on the horizontal direction, the LSTM units are difficult to model the local correlation of the original vertically adjacent patches because they are far apart in the sequence. Therefore, for this case, the local correlation in the CT image along the vertical direction will be corrupted.

Hilbert space-filling curve is a continuous non-conductible fractal curve that fills a flat square, which was first proposed by David Hilbert in 1891. The Hilbert curve has a significant advantage in maintaining local correlations when converting CT images from multi-dimensional to one-dimensional space compared to other space-filling curves [25]. Figure 5 illustrates the Hilbert space-filling curves for different orders (from left to right: first order, second order and third order, respectively).
the serialization sequence of $Z$ from two-dimensional to one-dimensional space based on the corresponding coordinates.

Figure 4 shows the overall structure of the LG-LSTM module. Formally, given the feature tensor $M \in \mathbb{R}^{H \times W \times C}$ as input to the LG-LSTM module, we first serialize $M$ based on the Hilbert curve-based serialization algorithm and obtain a new series of feature sequences $M' \in \mathbb{R}^{H' \times W' \times C}$; second, model the spatial relationship between the semantic patches of CT image based on $2HW$ LSTM units, and obtain the enhanced feature tensor $M'' \in \mathbb{R}^{H' \times W' \times C}$; and finally, we de-serialize and reshape $M''$ according to the serialized index to get the final enhanced feature map $\hat{M} \in \mathbb{R}^{H \times W \times C}$ as output to the LG-LSTM module.

The NMC module seamlessly connects the multi-scale attention mechanism with the LG-LSTM module, which can enable the pixel-level segmentation network to apply local guidance from the neighbourhood location and global guidance from the full map to each pixel. Such mechanism can better mine the complex short-range and long-range contextual information, thereby effectively improving the segmentation performance of CAC. Besides, the NMC module also effectively reduces the computational complexity of our AMSN. Note that the NMC module is simple and can be easily inserted into any segmentation backbones.

### 2.3 The hybrid loss for optimizing AMSN

After observation, it was found that intensity ambiguity usually makes the boundaries of CAC difficult to identify. Some previous literature also proved the importance of mining boundary information for semantic segmentation [26]. Typically, since the pixel-level cross-entropy loss function mainly uses pixel accuracy as a metric, for some difficult samples, mainly in the boundary region, the cross-entropy loss tends to generate an ambiguous response. In contrast, because the DSC loss can measure the overlapping part of the samples, it can better model the boundary information of CAC. Besides, the DSC loss can also solve the problem of uneven imbalance in the segmentation of CAC. However, since the traditional DSC loss will cause numerical instability in training, we rewrite it using Equation (5) to increase the numerical stability of the training process.

$$
L_{\text{seg}} = \frac{1}{C} \sum_{c=1}^{C} \sum_{i=1}^{HW} \log(\hat{y}_{ci} + \gamma_0 + 1) - \log(\hat{y}_{ci} \gamma_0 + 1) \quad (5)
$$

where $H$ and $W$ represent the width and height of the chest CT image, and $\gamma$ and $\hat{y}$ represent the reference segmentation and predicted segmentation, respectively. $C$ represents the number of class, which is set to 2. Note that by adding 1, i.e. Laplace smoothing, to the left and right terms, respectively, the division-zero problem similar to that of the original DSC loss can be avoided.

For our task, the small size of CAC and their close associations with the surrounding tissues made their boundaries difficult to distinguish. Therefore, we introduce a new boundary-preserving loss to further enhance the segmentation accuracy of CAC boundaries. Different from the methods [27, 28], which try to encode the distance from each pixel to its nearest pixel on the boundary line, we try to encode the average value of the distance from each pixel to the closest pixel on the boundary line of all CAC within CT image as follows:

$$
\hat{b}_i = \exp \left( -\frac{\sum_{m=1}^{3} \min_{c \in C} \| r_i - c'_j \|^2 }{2\delta_d^2} \right) \quad (6)
$$

where $r_i$ represents the coordinates of $i$th pixels, and $C$ represents the set of all pixels on the CAC boundary. Note that in such a way, not only the boundary line information is encoded, but also the relationship between different CAC is encoded. This is equivalent to the global optimization of the relationship between CAC, rather than a separate optimization.

Based on Equation (6), the boundary-preserving loss $L_{bp}$ can be formulated as follows:

$$
L_{bp} = \frac{1}{N} \sum_{i=1}^{N} \| \hat{b}_i - b_i \|^2_2 \quad (7)
$$

where $\hat{b}_i$ is the prediction result of the $i$th pixel and $N$ is the number of pixels of the CT image.

By combining the Laplace smoothing-based DSC loss and the boundary-preserving loss, we develop a new hybrid loss, which can not only reduce the imbalance of positive and negative samples but also capture fine structures with clear boundaries for CAC segmentation. The hybrid loss $L_{\text{total}}$ is defined as follows:

$$
L_{\text{total}} = L_{\text{seg}} + L_{bp} + \lambda \| \mathcal{W} \|^2_2 \quad (8)
$$

where $\lambda$ is the weight decay term and $\mathcal{W}$ is the parameters of the entire network.

### 3 EXPERIMENTS

In this section, we have done extensive experiments to evaluate the performance of our AMSN. First, we evaluate AMSN with five commonly used metrics and five existing state-of-the-art methods. Next, we conduct ablation studies on the newly proposed module and the hybrid loss. Then, a visual comparison between AMSN and existing state-of-the-art algorithms is conducted. Finally, we invited a radiologist to manually score the segmentation results of CAC, which further demonstrates the potential of AMSN in the clinical setting.

#### 3.1 Dataset

For training and evaluation, our study comprises the contrast-enhanced chest CT sequence of 130 CVD patients obtained from Qianfoshan Hospital in Shandong Province, where each patient contained tens of CT slices, and each DCM file...
represents one of the axial planes of each patient’s CT sequence. These dozens of DCM files make up the CT image scan sequence of one patient. The resolution of each CT scan is $512 \times 512$ pixels, with in-plane resolution ranging from 0.49 to 0.98 mm; the slice spacing ranging from 0.6 to 3.0 mm; the slice thickness ranging from 1 to 3 mm; and the tube voltage set to 120 or 140 kVp. Specifically, the dataset was manually annotated by two medical students, and have been revised by a medical specialist.

As shown in Figure 1, after the step of candidate region generation, we first use the method of literature [29] and the nearest neighbour interpolation to resize the roughly localized CT image and the label image to $256 \times 256$, respectively. Then, we perform a normalization operation for each CT due to the large differences in contrast between different CT images. For a given chest CT image, we normalize it based on the expectation and variance of the entire CVD dataset. It is worth noting that our training process is not based on any data augmentation operations.

### 3.2 Training protocols

All networks are implemented with PyTorch 1.1.0 in Python 2.7, with the batch size of 128, training epochs of 150 and learning rate of 0.1 (decayed by 10 every 2 epochs). All experiments are conducted on one Intel Xeon(R) E5-2620@2.5GHz and one Nvidia GeForce GTX TITAN X with CUDA v9.0, cuDNN v7.0.

### 3.3 Evaluation metrics

#### 3.3.1 Mean pixel accuracy

The mean pixel accuracy (MPA) is a simple metric that calculates the proportion of correctly classified pixels in each class and then averages the results, defined as

$$\text{MPA} = \frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} \frac{P_c}{P_i}$$

(9)

where $N$ is the number of samples, $C$ is the number of class, $P_c$ is the number of correctly classified pixels of class $c$ and $P_i$ is the number of pixels of class $i$.

#### 3.3.2 Intersection over union

The intersection over union (IoU) is a standard metric for semantic segmentation. Given a set of images, the IoU gives the similarity between the reference segmentation $S_{\text{ref}}$ and predicted segmentation $S_{\text{pred}}$, defined as

$$\text{IoU} = \frac{|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{pred}} \cup S_{\text{ref}}|}$$

(10)

#### 3.3.3 Dice similarity coefficient

The DSC is one of the commonly used metrics for evaluating the segmentation performance, which can be used to measure the similarity between two sets, defined as

$$\text{DSC} = \frac{2|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{pred}}| + |S_{\text{ref}}|}$$

(11)

#### 3.3.4 Recall rate

The recall rate is used to measure how many positive examples in the samples are predicted correctly, defined as

$$\text{Recall} = \frac{|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{ref}}|}$$

(12)

#### 3.3.5 Mean surface distance

We further evaluate the segmentation performance by measuring mean surface distance (MSD), defined as

$$\text{MSD} = \frac{1}{2} \left( \sum_{j=1}^{n_S} \min_{p' \in S'} \|p_j - p'_j\|_2 + \sum_{j'=1}^{n_S'} \min_{p_j \in S} \|p'_j - p_j\|_2 \right)$$

(13)

where $S$ and $S'$ are the ground-truth image and the predicted image, respectively.

### 3.4 Experimental design and results

In this section, five state-of-the-art semantic segmentation methods, FCN [30], SegNet [31], U-Net [32], DeepLab v3+ [33] and nnU-Net [34, 35], are trained on our CVD dataset, and compared with our AMSN and its various variants. All experiments are conducted under four-fold cross-validation, that is, the CVD dataset is randomly divided into 98 patients for training, and the remaining 32 patients for undiscovered testing. To be fair, all the above models are implemented on PyTorch 1.1.0 strictly following the original papers.

For evaluation, we conducted various experiments to verify the effectiveness of our AMSN. For objective evaluation: 1) We compare our AMSN with the existing state-of-the-art methods based on the above-mentioned evaluation metrics; 2) we conduct extensive ablation experiments on the CVD dataset to evaluate the effectiveness of each module of AMSN; 3) to prove the effectiveness of our proposed hybrid loss $\mathcal{L}_{\text{total}}$, we compared the performance of its ablation versions and the conventional cross-entropy loss; 4) we analyzed the processing time of our AMSN; 5) we verified the effect of different batch sizes on the convergence speed and the segmentation performance; and 6) we also verified the effectiveness of our candidate region generation method.
TABLE 1  Quantitative comparisons between our AMSN and the existing state-of-the-art methods in terms of MPA, Mean IoU, Recall, MSD and DSC under four-fold cross-validation

| Method         | MPA    | Mean IoU | Recall   | MSD    | DSC     |
|----------------|--------|----------|----------|--------|---------|
| FCN            | 88.33 ± 4.03 | 69.73 ± 4.32 | 81.28 ± 3.62 | 3.02 ± 0.87 | 64.93 ± 5.14 |
| SegNet         | 89.23 ± 3.82 | 71.02 ± 3.73 | 89.31 ± 3.58 | 2.97 ± 0.75 | 66.01 ± 4.57 |
| U-Net          | 90.03 ± 3.75 | 71.23 ± 3.55 | 80.64 ± 3.37 | 2.73 ± 0.62 | 70.38 ± 4.53 |
| DeepLab v3+    | 91.14 ± 4.17 | 74.28 ± 3.81 | 83.31 ± 3.48 | 2.62 ± 0.67 | 70.09 ± 4.72 |
| nnU-Net        | 91.83 ± 3.62 | 76.14 ± 3.37 | 89.24 ± 3.31 | 2.31 ± 0.55 | 71.43 ± 3.92 |
| AMSN           | 92.72 ± 3.71 | 77.81 ± 3.42 | 90.83 ± 3.27 | 2.03 ± 0.53 | 73.05 ± 4.16 |

Table 1 tabulates the MPA, IoU, DSC, Recall and MSD comparison values with standard deviations of AMSN and all existing state-of-the-art methods. It can be seen from the table that our AMSN has achieved the best results under all evaluation metrics. This not only proves the effectiveness and advantages of AMSN but also shows that our newly designed different modules have improved the segmentation performance of the model to different degrees. Specifically, AMSN outperforms all the listed methods by a large margin and finally achieves 92.72% MPA, 77.81% mean IoU, 90.83% recall, 2.03% MSD and 73.05% DSC, respectively.

For subjective evaluation: 1) We visually compare the results of AMSN with other state-of-the-art methods to further verify its effectiveness; and 2) we also conducted a qualitative assessment of the CAC segmentation results.

### 3.4.1 Objective evaluation

Table 1 tabulates the MPA, IoU, DSC, Recall and MSD comparison values with standard deviations of AMSN and all existing state-of-the-art methods. It can be seen from the table that our AMSN has achieved the best results under all evaluation metrics. This not only proves the effectiveness and advantages of AMSN but also shows that our newly designed different modules have improved the segmentation performance of the model to different degrees. Specifically, AMSN outperforms all the listed methods by a large margin and finally achieves 92.72% MPA, 77.81% mean IoU, 90.83% recall, 2.03% MSD and 73.05% DSC, respectively.

Table 2 verifies the performance of AMSN on the test set by adding different numbers of NMC modules. To facilitate testing, we regard the main branch as the baseline. We can observe that, adding NMC1 module to the baseline improves the performance by 1.88% DSC compared with the baseline, which demonstrates the effectiveness of the NMC module in multiscale feature fusion. Furthermore, to add NMC2 to the baseline can further improve the performance by 2.26% DSC, an increase of 0.38% than just adding NMC1 module, demonstrating that higher level semantic feature can better guide the NMC module for feature learning. Note that, based on the evaluation of the current CVD dataset, we only added two NMC modules to the main branch, in the hope that AMSN can balance the importance between segmentation accuracy and computational complexity to the greatest extent.

Table 2 tabulates the ablation studies results of different variants of the newly designed hybrid loss function. To facilitate discussion, we use D and B to denote $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{bp}}$ of Equation (8), respectively. It can be found that our proposed hybrid function (D + B + W) achieves 73.05% DSC accuracy. First, after removing the boundary loss (D + W), DSC is 72.53%, a decrease of 0.52%, which demonstrates the boundary loss effectively improves the segmentation accuracy of the network by considering the boundaries of all CAC as constraints. Then, after removing the regularization term (D + B), DSC is 72.92%, a decrease of 0.13%, which demonstrates by constraining the parameters of the entire network, not only reduces the probability of gradient explosion or gradient disappearance to a certain extent but also can make the optimized curve of the DSC loss more smooth. To further evaluate the effectiveness of AMSN, we also replace our newly designed hybrid loss with the conventional cross-entropy (CE) loss, which achieves 72.14% DSC accuracy, a decrease of 0.91% than the hybrid loss. Such comparison demonstrates that our proposed hybrid loss can

| Method       | MPA    | Mean IoU | Recall   | MSD    | DSC     |
|--------------|--------|----------|----------|--------|---------|
| MAIN         | 88.72 ± 3.92 | 73.63 ± 3.62 | 86.21 ± 3.15 | 2.85 ± 0.59 | 69.87 ± 4.45 |
| MAIN + NMC1  | 91.21 ± 4.01 | 76.21 ± 3.71 | 89.03 ± 3.26 | 2.23 ± 0.61 | 71.75 ± 4.51 |
| MAIN + NMC2  | 91.35 ± 3.63 | 77.01 ± 3.55 | 89.41 ± 3.31 | 2.17 ± 0.47 | 72.13 ± 4.25 |
| AMSN         | 92.72 ± 3.71 | 77.81 ± 3.42 | 90.83 ± 3.27 | 2.03 ± 0.53 | 73.05 ± 4.16 |

Table 2 tabulates the MPA, IoU, DSC, Recall and MSD comparison values with standard deviations of AMSN and all existing state-of-the-art methods. It can be seen from the table that our AMSN has achieved the best results under all evaluation metrics. This not only proves the effectiveness and advantages of AMSN but also shows that our newly designed different modules have improved the segmentation performance of the model to different degrees. Specifically, AMSN outperforms all the listed methods by a large margin and finally achieves 92.72% MPA, 77.81% mean IoU, 90.83% recall, 2.03% MSD and 73.05% DSC, respectively.

Table 3 tabulates the ablation studies results of different variants of the newly designed hybrid loss function. To facilitate discussion, we use D and B to denote $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{bp}}$ of Equation (8), respectively. It can be found that our proposed hybrid function (D + B + W) achieves 73.05% DSC accuracy. First, after removing the boundary loss (D + W), DSC is 72.53%, a decrease of 0.52%, which demonstrates the boundary loss effectively improves the segmentation accuracy of the network by considering the boundaries of all CAC as constraints. Then, after removing the regularization term (D + B), DSC is 72.92%, a decrease of 0.13%, which demonstrates by constraining the parameters of the entire network, not only reduces the probability of gradient explosion or gradient disappearance to a certain extent but also can make the optimized curve of the DSC loss more smooth. To further evaluate the effectiveness of AMSN, we also replace our newly designed hybrid loss with the conventional cross-entropy (CE) loss, which achieves 72.14% DSC accuracy, a decrease of 0.91% than the hybrid loss. Such comparison demonstrates that our proposed hybrid loss can

| DSC (%) | CE    | D + W | D + B | D + B + W |
|---------|-------|-------|-------|-----------|
| Mean    | 72.14 | 72.53 | 72.92 | 73.05     |
| Std     | 3.91  | 4.23  | 4.03  | 4.16      |
| Min     | 51.46 | 52.31 | 53.74 | 54.11     |
| Max     | 82.71 | 84.97 | 85.53 | 85.65     |
significantly improve the performance of the training model, which in some ways also proves that the traditional cross-entropy loss does not deal well with the sample imbalance problem. Therefore, our proposed hybrid loss makes AMSN an efficient and reliable clinical segmentation tool for CAC segmentation by comprehensively utilizing the advantages of the boundary-preserving loss and the Laplace smoothing-based DSC loss.

To explain the processing time of our AMSN, we conducted experiments on its parameters amount and the inference time. Note that the parameters amount is related to the backward training time of the model, and the inference time is related to the forward testing time of the model. Obviously, such two times indicators are most related to the processing time, which can further demonstrate the effectiveness of our AMSN. As shown in Table 4, we can see that our NMC module has fewer parameters, which proves that our AMSN has less computational complexity in the reverse training process. Second, our NMC module also has less test time, which, on the other hand, proves that our AMSN has a faster inference speed.

Figure 6 shows the training loss curve of our model when using different batch sizes. It is not difficult to find that different batch sizes can affect the smoothness of the loss curve but also affect the convergence time of the model. By comparing the performance of different batch sizes selected, we find that when the batch size is set to 128 (grey curve), a relatively optimal loss curve can be obtained. This demonstrates that different batch sizes can affect the segmentation performance of AMSN to a certain extent.

Besides, to further explain the effectiveness of our candidate region generation method, we also compare the mean IoU of the generated proposal area. Since we directly crop the candidate area (128 × 128) based on the barycentric coordinate of the proposal area, and the targets contained in the target area are all extremely small targets, we only need to compare the mean IoU of the proposal area with the maximum confidence to demonstrate the usefulness of our proposed candidate region generation method. Because as long as the distance between the barycentric coordinate of the proposal area and the target area is less than $\frac{128}{2}$, the cropped area (128 × 128) based on the barycentric coordinate of the proposal area can cover all CAC much more accurately. As shown in Table 5, the mean IoU values of all CT images are all above 80%, which demonstrates that 64 is much larger than the distance between the barycentric coordinate of the proposal area and the target area, so the cropped candidate regions can cover all CAC with nearly 100% recall among almost all used cases.

| TABLE 4 | Quantitative comparisons between our AMSN with the NMC module and its different versions with the original attention (OA) module and the dual attention (DA) module in terms of the parameters amount and the inference time |

| Setup | With 2 × OA [23] | With 2 × DA [11] | with 2 × NMC |
|-------|-------------------|-----------------|-------------|
| Params (M) | Forward and backward size | 141.28 | 164.28 | 97.31 |
| | Estimated total size | 138.95 | 160.16 | 93.78 |
| Test time (s/img) | Inference speed | 1.21 | 1.48 | 0.52 |

| TABLE 5 | Detection accuracy of the candidate generation method for the six randomly selected CT images in Figure 7 and the test set in terms of IoU |

| CT image | CT1 | CT2 | CT3 | CT4 | CT5 | CT6 | Test set |
|----------|-----|-----|-----|-----|-----|-----|----------|
| mIoU (%) | 90.3 | 92.6 | 85.2 | 91.7 | 83.2 | 84.0 | 85.32 |
3.4.2 Subjective evaluation

It should be noted that the segmentation performance of methods with higher objective evaluation values may not be always good. Therefore, to further test the effectiveness of our proposed method, we also visually compare the results of AMSN with other state-of-the-art methods. As shown in Figure 7, we selected six sets of test CT images, including two sets of small targets, two sets of targets and two sets of large targets. Overall, our AMSN has achieved the best visual effect in all cases, which proves that it has good generalization ability in different situations. Specifically, as we can see from the first and second subgraphs, there are false-positive samples of varying degrees in almost all other methods. Thanks to the LG-LSTM module, by modelling the path-wise features of the CT image, false-positive samples are effectively suppressed. It can be seen from the third, fifth and sixth subgraphs that the other methods do not perform well on boundary segmentation due to the blurred CAC boundaries and large shape variation, resulting in many false-positive samples. Thanks to the NMC module and hybrid loss function effectively maintaining the boundary, our AMSN can
not only obtain reliable boundaries but also effectively suppress most false-positive samples.

We also propose a qualitative evaluation algorithm to verify the effectiveness of AMSN. As shown in Figure 8, an expert uses the quality scoring method mentioned in [36] to visually inspect and grade the segmentation results of CAC. In the clinical setting, we believe that such excellent results paved a good way for AMSN as an accurate CAC segmentation tool for assisted CVD diagnosis.

4 CONCLUSION

This paper proposes AMSN, a new two-branch network for accurate CAC segmentation on chest CT images. Given that too much background can interfere with CAC segmentation, we first use morphological manipulations based on the prior knowledge to obtain over-segmented results of the coronary artery region rather than directly using the original CT images for segmentation. Then, we introduce a novel NMC, which uses high-level features as boundary constraints to enhance the representational capability of low-level features, allowing the network to generate more accurate and clear boundaries and significantly improve the segmentation accuracy of CAC. Finally, we also introduced a new hybrid loss to further improve the stability and segmentation accuracy of our model. Experiments show that our AMSN achieves state-of-the-art results on the newly proposed CVD dataset. In the clinical setting, such advantages are greatly needed.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (No. 61572292, 61332015, and 61602277), and in part by the NSFC Joint Fund with Zhejiang Integration of Informatization and Industrialization (No. U1609218).

REFERENCES

1. National Center for Health Statistics. Health, United States, 2016, With Chartbook on Long-Term Trends in Health. Number 2017. Government Printing Office, USA (2017)
2. Lessmann, N., et al.: Automatic calcium scoring in low-dose chest ct using deep neural networks with dilated convolutions. IEEE Trans. Med. Imaging 37(2), 615–625 (2017)
3. Zhang, W., et al.: An end-to-end joint learning framework of artery-specific coronary calcium scoring in non-contrast cardiac ct. Computing 101(6), 667–678 (2019)
4. Wu, J., et al.: A fully automated multi-modal computer aided diagnosis approach to coronary calcium scoring of MSCT images. In: Medical Imaging 2012: Computer-Aided Diagnosis, vol. 8315, p. 83152I. International Society for Optics and Photonics, USA (2012)
5. Ding, X., et al.: Automated coronary artery calcium scoring from non-contrast CT using a patient-specific algorithm. In: Medical Imaging 2015: Image Processing, vol. 9413, p. 94132U. International Society for Optics and Photonics, USA (2015)
6. Eliot, D., Goldenberg, R.: Fully automatic model-based calcium segmentation and scoring in coronary ct angiography. Int. J. Comput. Assisted Radiol. Surg. 9(4), 595–608 (2014)
7. Ahmed, W., et al.: Automatic detection and quantification of the Agatston coronary artery calcium score on contrast computed tomography angiography. Int. J. Cardiovascular Imag. 31(1), 151–161 (2015)
8. Wolterink, J.M., et al.: Automatic coronary calcium scoring in non-contrast-enhanced ecg-triggered cardiac CT with ambiguity detection. IEEE Trans. Med. Imaging 34(9), 1867–1878 (2015)
9. Wolterink, J.M., et al.: Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med. Image Anal. 34, 123–136 (2016)
10. Yuan, Y., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916, 2018
11. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154. IEEE, USA (2019)
12. Huang, Z., et al.: Cenet: Cross-context attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 603–612 (2019)
13. Ren, S., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99. Curran Associates, Inc., USA (2015)
14. Chen, L.-C., et al.: Rethinking Atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05887 (2017)
15. Ning, Y., et al.: Automated pancreas segmentation using recurrent adversarial learning. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 927–934 (2018)
16. Ning, Y., et al.: DRAN: Deep recurrent adversarial network for automated pancreas segmentation. IET Image Proc. 14(6), 1091–1100 (2019)
17. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
18. Zhang, T., et al.: Decoupled spatial neural attention for weakly supervised semantic segmentation. IEEE Trans. Multimedia 21(11), 2930–2941 (2019)
19. Li, X., et al.: GFF: Gated fully fusion for semantic segmentation. arXiv preprint arXiv:1904.01803 (2019)
20. Tang, Y., et al.: Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE Trans. Image Process. 28(10), 4997–5012 (2019)
21. Zhang, X., et al.: Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722. IEEE, USA (2018)
22. Zhang, Y., et al.: Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082 (2019)
23. Ren, S., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 6000–6010 (2017)
24. Breinholst, G., Schierz, C.: Algorithm 781: Generating Hilbert’s space-filling curve by recursion. ACM Trans. Math. Software 24(2), 184–189 (1998)
25. Moon, B., et al.: Analysis of the clustering properties of the Hilbert space-filling curve. IEEE Trans. Knowl. Data Eng. 13(1), 124–141 (2001)
26. Chen, H., et al.: DCAN: Deep contour-aware networks for object instance segmentation from histology images. Med. Image Anal. 36, 135–146 (2017)
27. He, K., et al.: Pelvic organ segmentation using distinctive curve guided fully convolutional networks. IEEE Trans. Med. Imaging 38(2), 585–595 (2019)
28. Pfister, T., et al.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1913–1921 (2015)
29. Ning, Y., et al.: Adaptive image rational upsampling with local structure as constraints. Multimedia Tools Appl. 78(6), 6889–6911 (2019)
30. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
31. Badinarayanan, V., et al.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2017)
32. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Germany, (2015)
33. Chen, L.-C., et al.: Encoder-decoder with Atrous separable convolutions for semantic image segmentation. In: Proceedings of the European
conference on computer vision (ECCV), pp. 801–818. Springer International Publishing, Germany (2018)

34. Isensee, F., et al.: NNU-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)

35. Isensee, F., et al.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2020)

36. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88 (2017)

How to cite this article: Ning Y, Zhang Y, Li X, Zhang C. Towards accurate coronary artery calcium segmentation with multi-scale attention mechanism. IET Image Process. 2021;15:1359–1370.
https://doi.org/10.1049/ipr2.12110