3D U-Net With Attention and Focal Loss for Coronary Tree Segmentation

Riming Liang

The First People's Hospital of Foshan

xingsheng huang (✉ xingsheng.huang@raysightmed.com)

Shenzhen Raysightmed Intelligent Medical Technology Co. Ltd. <https://orcid.org/0000-0001-7647-944X>

Jun Ma

Shenzhen Raysightmed Intelligent Medical Technology Co. Ltd.

Guomin Ma

LiaoNing Provincial People's Hospital

Kunhua Wang

LiaoNing Provincial People's Hospital

Research Article

Keywords: Medical Image Segmentation, Coronary tree, U-Net, Channel attention, Spatial attention, focal loss

Posted Date: July 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-717054/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

3D U-Net with attention and focal loss for coronary tree segmentation

Riming Liang^{1#}, Xingsheng Huang^{2#}, Jun Ma^{2#}, Guomin Ma³, KunHua Wang^{3*}

¹The First People's Hospital of Foshan

²Shenzhen Raysightmed Intelligent Medical Technology Co., Ltd., Shenzhen 518000, People's Republic of China

³LiaoNing Provincial People's Hospital

63873820@qq.com,
{xingsheng.huang, jun.ma}@raysightmed.com,
2638785457@qq.com, wang-christina@163.com

Abstract. The semantic segmentation of coronary artery is important in clinical diagnosis and treatment of coronary artery disease (CAD). The problems of intra-class inconsistency and inter-class indistinction in coronary artery and vein are very prominent. In this paper, we propose a 3D U-Net model improvement plan using attention mechanism and focal loss. U-Net is combined with channel attention and spatial attention to distinguish confusing categories and targets with similar appearance features. We use focal loss to optimize the loss function to solve the problem of category imbalance between coronary artery and background category.

Keywords: Medical Image Segmentation · Coronary tree · U-Net · Channel attention · Spatial attention · focal loss

1 Introduction

Coronary artery disease (CAD) is the leading cause of death worldwide, constituting more than one-fourth of global mortalities every year [1]. The semantic segmentation of coronary arteries is very important in the clinical practice of CAD diagnosis and treatment. First, as morphological evidence, precise coronary artery segmentation can preview the topology of the coronary artery tree, and help radiologists accurately measure stenosis, and further analyze the plaque lesions in the stenosis. Secondly, accurate coronary artery segmentation is the foundation for functional analysis such as non-invasive CT-based fractional flow reserve (CTFFR).

Convolutional neural networks have been widely applied in the field of medical image segmentation. One of the most used models is U-Net, which is proposed in 2015 by Olaf Ronneberger et al. [2] U-Net is a semantic segmentation network based on encoder-decoder, suitable for the segmentation of medical images. The U-Net network structure includes a down-sampling stage (encoder) and an up-sampling stage (decoder), each of which consists of several convolutional layers and pooling layers, with no fully connected layers. The shallow features in the network are used to generate lower-level features and the deep features are used to generate higher level features, so that the comprehensive image semantic level segmentation can be realized. The up-sampling stage and the down-sampling stage in U-Net use the same number of levels of convolution operations, and the skip connection structure is used to connect the down-sampling layer and the up-sampling layer. The skip

Riming Liang, Xingsheng Huang and Jun Ma contributed equally to this work

* Corresponding author

connection operation transfers features extracted by the down-sampling layer directly to the up-sampling layer, which makes pixel positioning more accurate and segmentation accuracy higher.

In the task of segmentation of the coronary artery tree, we focus on the characteristics of the coronary tree's fine lumen and rich topology. However, in the actual coronary tree segmentation work, 3D U-Net [3] model is difficult to distinguish between confusing categories and targets with similar appearance features. For example, coronary arteries and veins have very similar visual features in topology, shape and brightness. In response to this problem in coronary tree segmentation, the dual attention model [4] captures rich context dependencies based on a self-restraint mechanism, which can effectively solve the problems of intra-class inconsistency and inter-class indistinction in segmentation tasks.

In this paper, we propose an improvement plan for the 3D U-Net model. First, we incorporated channel attention, spatial attention, and a combination of the two to make use of the relationship between coronary tree targets. Secondly, we optimized the loss calculation during training.

According to the main remarks highlighted above, we propose an improved 3D U-Net model to optimize the segmentation of the coronary artery tree. In particular, the main contributions of our work are:

- The improved attention models enrich the details of coronary tree segmentation.
- The focal loss based on the 3D U-Net model pays more attention to the foreground category and participates in the gradient calculation.
- Proper isotropic spacing parameters improve the results of coronary artery tree segmentation.
- Better segmentation performance has been achieved by the proposed model.

The rest of this paper is organized as followings. The second part briefly introduces the related work for image segmentation. The third part details the architecture of our proposed model. The fourth part shows the experiments setting and results. Finally, the conclusion and future work are drawn in section five.

2 Related work

Segmentation of coronary with CNN

Quantitative analysis of coronary arteries is an important step for the diagnosis of cardiovascular diseases, stenosis grading, blood flow simulation and surgical planning. With respect to the coronary artery segmentation, most deep learning based approaches use an end-to-end CNN segmentation scheme to predict dense segmentation probability maps [5-8]. In particular, Moeskops et al. [5] proposed a multi-task segmentation framework where a single CNN can be trained to perform three different tasks including coronary artery segmentation in cardiac CTA and tissue segmentation in brain MR images. They showed that such a multi-task segmentation network in multiple modalities can achieve equivalent performance as a single task network.

Besides, shape priors can also be incorporated into the convolutional neural network [9-11]. For instance, Lee et al. [9] explicitly enforced a roughly tubular shape prior for the coronary segments by introducing a template transformer network, through which a shape template can be deformed via network-based registration to produce an accurate segmentation of the input image, as well as to guarantee topological constraints. Lee et al. [9] showed that such method significantly outperformed a baseline network that used only fully-connected layers on healthy subjects (mean Dice score: 0.75 vs. 0.67).

U-Net with attention

Oktay et al.[12] proposed an attention gating (AG) model for medical imaging, which can automatically focus on target structures of different shapes and sizes. Specifically, the model trained with AG can suppress irrelevant areas in the input image while highlighting salient features useful for specific tasks. This eliminates the need to use cascaded convolutional neural networks (CNN) and cascade tissue/organ positioning modules. AG can be easily integrated into standard CNN architectures, such as U-Net models, with small computational overhead, while improving model sensitivity and prediction accuracy.

Xu et al. [13] proposed the first visual attention model in image captioning. Usually, they used “hard” pooling to select the most likely attentive area, or use “soft” pooling to average spatial features and attentive weights. As for VQA, Zhu et al. [14] use “soft” attention to merge image area features. To further improve spatial attention, Yang et al. [15] applied a stacked spatial attention model, where the second attention is based on the attentive feature map modulated by the first one. Different from theirs, Chen et al. [16] multi-layer attention is applied on the multiple layers of a CNN. A common defect of the above spatial models is that they generally resort to weighted pooling on the attentive feature map. Thus, spatial information will be lost inevitably [16]. More seriously, their attention is only applied in the last conv-layer, where the size of receptive field will be quite large and the differences between each receptive field region are quite limited, resulting in insignificant spatial attentions [16].

Tolooshams et al. [17] proposed an end-to-end neural architecture for multichannel speech enhancement, call Channel-Attention Dense U-Net. The distinguishing feature of the proposed framework is a channel attention (CA) mechanism inspired by beamforming. CA is motivated by the self-attention mechanism, which captures global dependencies within the data. This paper incorporates CA into a U-Net to guide the network to decide, at every layer, which feature maps to pay the most attention to.

Vessel segmentation based on U-Net

Chen et al. [18] proposed to incorporate the vesselness map into the input of the 3D U-Net, which serves as the reinforced information to highlight the tubular structure of coronary arteries. In the proposed architecture, the input is composed of two channels of the same volume of interest, one from the original CTA image and the other from the vesselness map derived by applying Frangi filtering to the original CTA image. This study included 33472 training samples (11 cases), 5683 and 12223 validation samples(2 cases), 6841 and 7028 testing samples(2 cases), it achieved Dice coefficient of

0.806 by applying the largest connected component after the output of the segmentation network.

Livne et al. [19] proposed the half U-Net, where the number of channels in each layer was consistently reduced to half. The half U-Net was fed with cerebrovascular 2D image patches and returned the 2D segmentation probability map for each given patch. The half U-net model was trained on 81,000 extracted and augmented patches from 41 patients, validated using 11 full patient volumes and assessed for performance using the test-set of 14 full patient volumes. The half U-net model yielded a Dice result of 0.89.

Kong et al. [20] proposed a novel 3D tree-structured convolutional gated recurrent unit (TreeConvGRU) model to learn the anatomical structure of the coronary artery. Kong et al. collected four large datasets (916 CT scans in total) from four hospitals, where 80%, 5%, and 15% scans were used for training, validation, and testing, respectively. Based on the results of pre-segmentation of coronary arteries using a 3D U-Net network, the average Dice score performance of the TreeConvGRU model for coronary optimization reached 0.8683, which was an improvement of 0.71% over ConvGRU.

CNN optimization with loss function

Lin et al. [21] proposed the focal loss idea for the problem of sample imbalance, mainly to solve the problem of serious imbalance of positive and negative sample ratio in one-stage target detection. This loss function reduces the weight of a large number of simple negative samples in training, and can also be understood as a kind of difficult sample mining. State-of-the-art accuracy and running time are achieved on the challenging COCO data set.

3 Method

The 3D U-Net [3] model is one of the most widely used approaches in medical image segmentation. It has a typical encoder-decoder structure, as shown in Figure 1.

Compared with ASPP [22], PSPNet [23], LargeKernel [24] and other models that integrate multi-scale contextual information, the encoder-decoder structure of 3D U-Net better integrates the semantic information of the low, middle and high levels. Our 3D U-Net focuses on the coronary artery tree segmentation task to capture characteristics of the coronary tree lumen and rich topology.

Fig.1. The 3D U-Net architecture. Blue boxes represent feature maps. The number of channels is denoted above each feature map.

However, in the actual coronary artery tree segmentation work, it is challenging for the 3D U-Net model to distinguish confusing categories and targets with similar appearance features. For example, coronary arteries and veins are similar in topology, shape and brightness. It is necessary to strengthen the feature expression of intra-class inconsistency and inter-class indistinction for 3D U-Net models. Therefore, we propose a U-Net combined with attention mechanism to use the relationship between coronary tree targets.

3.1 U-Net with channel attention

Channel attention module aims to capture the interdependence between channels, and adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels.

Fig.2. Channel attention module architecture

The structure of channel attention module is illustrated in Figure 2. To make the description simpler, we use 2D CNN to illustrate the process, and 3D cases can be extended accordingly. The input

$\mathbf{A}$ is a collection of 2D feature maps with height H , width W and channel number C . Each 2D feature map of $\mathbf{A}$ is flattened to a vector of length $N=H \times W$, thus a matrix $\tilde{\mathbf{A}}$ of size $C \times N$ is formed. Multiplying $\tilde{\mathbf{A}}$ with its transpose and normalizing the result with softmax, one can obtain the map $\mathbf{X}$ ($C \times C$), which represents the pair-wise interdependency for all C channels.

$$X_{ji} = \frac{e^{A_i \cdot \tilde{A}_j}}{\sum_{i=1}^C e^{A_i \cdot \tilde{A}_j}}$$

After multiplying $\tilde{\mathbf{A}}$ with $\mathbf{X}$, the output is further scaled by a factor β , and then added to the original feature map $\mathbf{A}$ to obtain the final output $\mathbf{E}$.

$$\mathbf{E}_j = \beta \sum_{i=1}^C X_{ji} \mathbf{A}_i + \tilde{\mathbf{A}}_j$$

More detailed explanation of channel-wise attention is as follows. The $\mathbf{X}$ matrix plays the role of attention, where value $\mathbf{X}(i,j)$ represents the dependency between channel i and channel j . The matrix value is normalized by softmax to the range of $[0, 1]$, where larger values indicate stronger dependence. Multiplying this attention with $\tilde{\mathbf{A}}$ implies selectively integrating highly dependent channels, and hence improving semantic feature expression. Thus, semantic dependence between channels is modeled and feature maps are re-calibrated with this dependence.

An improved U-Net model (U-Net with CAM) combining U-Net with the channel attention module is proposed. We added the channel attention module before the last convolutional layer of the decoder to help U-Net with CAM pay more attention to contributions between channels. The specific network structure of U-Net with CAM is shown in Figure 3.

Fig.3. U-Net with CAM architecture

3.2 U-Net with spatial attention

Spatial attention module (SAM) is designed to capture the spatial dependence between any two positions of the feature map. For a particular position, its feature is updated according to similarities with features at all positions. Therefore, the positions with similar features can contribute to each other's improvement, regardless of the distance between them.

Context is important for medical image segmentation and it aims to capture global dependencies

regardless of spatial position. In order to model richer local context dependence, our method employs a SAM structure (as shown in Figure 4), encoding a wider context dependence to local features.

Fig.4. Spatial attention module architecture

As before, we use 2D scenario to describe the network structure. Input **A** ($C \times H \times W$) goes through three convolutional layers to obtain three feature maps of **B**, **C**, and **D**, which are reshaped to $C \times N$, where each of the **C** feature map is flattened to be a feature vector of length $N = H \times W$. Multiplying **B**'s transpose with **C** and followed by softmax, one gets the spatial attention map **S** ($N \times N$). Then, the product of **S** and all feature vectors of **D**, scaled by a factor, gives the final output map **E**.

Each element of the **S** matrix is:

$$S_{ji} = \frac{e^{B_i \cdot C_j}}{\sum_{i=1}^N e^{B_i \cdot C_j}}$$

Here are some remarks for spatial attention map **S**. The element S_{ji} represents the influence of pixel i on pixel j . Larger values indicate stronger relative dependence. Note, when calculating **S**, value at a pixel is the inner product on the channel dimension, not a single channel.

Each element of the final output **E** matrix is:

$$E_j = \alpha \sum_{i=1}^N S_{ji} D_i + A_j$$

Where α is the scale factor, D_i is the element of **D**, and A_j is the element of **A**. The attention is multiplied by the original map, i.e., the feature map is updated using the weighted sum of all positions, and the features is selectively strengthened according to similarities between pixels. It is equivalent to using the learned long-distance dependency.

Based on the above analysis of the spatial attention module, an improved U-Net model (U-Net with SAM) is proposed. We added the spatial attention module before the last convolutional layer of the decoder to help U-Net with SAM pay better attention to contributions between pixels. The specific network structure of U-Net with SAM is shown in Figure 5.

However, the map of spatial attention described above is very large in 3D, causing memory overflow. So, the original design of the spatial attention mechanism is not feasible for 3D, especially when image size is large. Therefore, we simplify it as shown in Figure 5. This spatial attention is obtained by 1×1 convolution to collapse the $C=32$ channels to one map of size $D \times H \times W$, and each channel of the feature maps shares the same spatial attention. This way, we can effectively reduce the use of memory space. In this setting, the spatial attention map does not represent the pixel-wise dependency, but it still guides the attention of the network to where it needs.

Fig.5. 3D U-Net with SAM architecture

3.3 U-Net with dual attention

In response to the aforementioned CAM and SAM, we also proposed a segmentation architecture that combines U-Net with CAM and SAM, namely U-Net with DAM, to enhance the discriminative ability of the feature representation of coronary tree segmentation.

Fig.6. 3D U-Net with DAM architecture

As shown in Figure 6, the feature map output by the penultimate layer of convolution passes through CAM and SAM, and then the features of CAM and the features of SAM are subjected to the sum operation, that is, “Sum fusion” in Figure 6, and the segmentation result of the final coronary

artery tree is obtained.

3.4 U-Net with loss optimization

In the coronary artery tree segmentation task, there is an extreme imbalance between the foreground (coronary) and background (non-coronary) categories. If we use the native cross entropy loss, the background dominates the loss, and it leads to low sensitivity. Therefore, based on previous work of [21], we use the focal loss to weigh the balance of foreground and background classes in semantic segmentation. The definition of Focal loss is:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

Where, p_t is

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$

According to conclusion of [21], $\gamma = 2$ gives the best performance. Therefore, in the coronary artery tree segmentation task, we mainly optimize for α with fixed $\gamma = 2$.

4 Experiments

In this part, we first introduce training and testing setup in detail in Section 4.1. In Section 4.2, we evaluate the segmentation performance of the coronary artery tree based on the U-Net with improved attention model. Then, we evaluate the impact of loss optimization on coronary segmentation performance in section 4.3. Finally, we test the effect of isotropic spacing on the performance of coronary artery tree segmentation in Section 4.3.

4.1 Data and Setting

In the experiments, we collected 300 cases of coronary artery CTA, all from the retrospective data of Liaoning Provincial People's Hospital. The 300 cases include patients with suspected coronary heart disease (coronary artery stenosis $>50\%$) and patients without significant stenosis (coronary artery stenosis $\leq 50\%$). Among them, plaques shown in coronary CT include calcified plaques, mixed plaques and non-calcified plaques. The coronary artery tree of each example is manually segmented and reviewed by radiologists with rich clinical experience, and the manual segmentations are used as ground truth for evaluation. The segmented coronary tree vessels include LM, LAD, LCX, RCA, D1, D2, D3, OM1, OM2, OM3, RI, PDA, AM1 and other blood vessels according to the AHA naming convention (17 paragraphs). On average, each example has 25.6 blood vessels.

We have samples with three categories according to coronary artery stenosis rates $>70\%$, $50-70\%$

and <50%, and labeled coronary plaque lesions. 200 cases are used as training set, and the remaining 100 cases are test set. The original spacing of each example is $0.6 \times 0.6 \times 0.5$ mm. In the experiment, it is resized to $0.6 \times 0.6 \times 0.6$ mm to achieve isotropy. The training input patch size is $128 \times 128 \times 128$, and the output is of the same size.

During training, all hyper-parameters follow those in 3D U-Net [3]. Specially, the initial learning rate is 0.001. The weight decay is 0.0001 and momentum is 0.98. In our experiments, when training epoch is set to be 80, and a total of 40k iterations, all the deep learning models are already converged. We adopt standard data augmentation methods and train the networks using SGD with a mini-batch size of 2 for each GPU. It takes 36 hours to train a model on NVIDIA GTX 1080Ti.

In the performance evaluation, we use the mean of class-wise intersection over union (Mean IoU), false positive, sensitivity and specificity to quantify the overall statistical performance of the coronary artery tree segmentation. Among them, mean IOU is measured by the voxel overlap, and it may cause bias towards thicker vessels because they contain more voxels. To balance this effect, sensitivity and specificity are measured by the length of coronary vessels centerlines, which are extracted by the skeleton operation.

4.2 Attention module for coronary segmentation

In this section, we try to conduct an in-depth discussion on the performance of 3D U-Net with CAM, SAM, and DAM models on coronary tree segmentation.

Table 1. U-Net with CAM, SAM and DAM model for coronary artery tree segmentation performance comparison

Model	Mean IoU	False Positive	Sensitivity	Specificity
3D U-Net with SAM	0.80138	0.19550	0.87677	0.90094
3D U-Net with CAM	0.80256	0.24336	0.90247	0.85851
3D U-Net with DAM	0.78092	0.18079	0.87069	0.89290
3D U-Net	0.80259	0.16986	0.87681	0.90402

For the coronary tree segmentation task, the segmentation performance of the 3D U-Net model is used as a benchmark. As seen from Table 1, for the mean IoU, the segmentation performance difference between 3D U-Net with CAM and U-Net with SAM is very small, and only U-Net with DAM has dropped by nearly 2%. For sensitivity, 3D U-Net has similar performance as U-Net with SAM and U-Net with DAM models, while the 3D U-Net with CAM model has a significant increase of more than 2%. Compared with manual labeled ground truth, the improvement of sensitivity is reflected by more branches or longer length of the coronary tree. Similarly, for specificity, 3D U-Net produce results close to U-Net with SAM and U-Net with DAM models, while the 3D U-Net with CAM model drops significantly by more than 4%. The main reason is that the 3D U-Net with CAM model captures

more and longer branches. Comparing with ground truth, the additional mismatched branch sections fall in two categories: (1) correct arteries but not segmented by hand in ground truth (2) wrong vessels. Both cases cause the specificity to decrease due to the significant increase in the denominator during calculation. However, we found that more cases belong to the first category than the second.

We show 4 very representative examples in Figure 7, with each row showing the result of an example from four methods. The example given includes model segmentation of stenosis on major coronary arteries (LAD, LCX, RCA). From the observation of the segmentation results, we find that all four methods can capture the stenosis of these four examples. These specific stenoses can be seen in the first row on the proximal LAD, the second row on the mid LCX, the third row on the mid LAD and distal RCA, and the fourth row on the proximal LAD. These specific stenoses segmentation situations can be seen in Figure 8. Those examples show that although the segmentation results from four methods are generally similar, U-Net with CAM has some advantage on completeness and correctness of the coronary tree. In the first row, U-Net with CAM correctly captures the two branches in the proximal-mid part of RCA, but other models miss one or both. In the second row, all three other methods show obvious false positives in the segmentation results, where veins near to the distal LAD are connected by mistake. U-Net with CAM avoided this wrong connection. The third and fourth row both show that U-Net with CAM captures branches that are missed by other methods.

Fig.7. Comparison of segmentation effects of U-Net with CAM, U-Net with SAM and U-Net with DAM models on four coronary tree examples.

One can see that U-Net with CAM has the best coverage of branches among the four methods, and also yields clean segmentation at the distal part. These observations are consistent with the statistical criteria summarized in the Table 1.

In summary, Mean IoU, False positive, sensitivity and specificity reflect the segmentation performance of the methods, and are consistent with the advantages and disadvantages of the segmentation visual effect for most examples.

Fig.8. The first row is the segmentation result of the U-Net with CAM model, and the second row is the curved planar reconstruction of the image corresponding to the coronary. The first row to the fifth row are the first row on the proximal LAD, the second row on the mid LCX, the third row on the mid LAD and distal RCA, and the fourth row on the proximal LAD in Fig. 7.

4.3 Loss for coronary segmentation

The results with focal loss and cross entropy loss (baseline) are shown in Table 2. Compared with baseline of U-Net model, one can find that the segmentation performance with focal loss has a significant decrease in the mean IoU, which is mainly due to the significant increase in false positives.

Although these false positives contain wrong segmentations, more of them are correct coronaries but not labeled by hand. Regarding the focal loss itself, we can find that when $\alpha=2$, gives better results than $\alpha=1$.

Table 2. Comparison of the segmentation performance of the coronary artery tree when the U-Net model is trained with different losses

Loss type	Mean IoU	False Positive	Sensitivity	Specificity
Focal loss($\alpha=1$)	0.71108	0.70545	0.94661	0.72229
Focal loss($\alpha=2$)	0.73283	0.63561	0.96083	0.73868
BCE (baseline)	0.80259	0.16986	0.87681	0.90402

4.4 Isotropic spacing

We also investigate the performance of four methods with different isotropic spacing. Specifically, the input to the networks are image patches of size 128 x 128 x 128, but we can choose different values of spacing for the isotropic resampling. The results are shown in Table 3.

Table 3. Comparison of the coronary tree segmentation performance of 3D U-Net, U-Net with CAM, U-Net with SAM and U-Net with DAM models under different spacing

Model	Spacing	Mean IoU	False Positive	Sensitivity	Specificity
3D U-Net with SAM	0.3	0.78579	0.32574	0.90412	0.80886
	0.6	0.80138	0.19550	0.87677	0.90094
	0.9	0.73808	0.16096	0.77837	0.93231
3D U-Net with CAM	0.3	0.77306	0.30986	0.86901	0.81946
	0.6	0.80256	0.24336	0.90247	0.85851
	0.9	0.76849	0.15098	0.81608	0.94374
3D U-Net with DAM	0.3	0.76097	0.37196	0.88198	0.78243
	0.6	0.78092	0.18079	0.87069	0.89290
	0.9	0.77823	0.16666	0.84114	0.93199
3D U-Net	0.3	0.80673	0.26553	0.86457	0.89011
	0.6	0.80259	0.16986	0.87681	0.90402
	0.9	0.77315	0.12329	0.81830	0.96372

Resampling spacing during training can change the field-of-view (FOV) of the model. Given the same input patch size, larger resampling spacing indicates larger FOV. When the FOV is large, the model has better modeling of the higher level information, like the overall shape and anatomy of the

coronary tree, but may lose some of the details. On the contrary, when the FOV is small, the details of the object can be learned well. However, it confuses true coronary arteries and other types of vessels. This tendency is shown in the table. The models with smaller spacing has higher sensitivity (indicating more vessels are captured), but the false positives are also high (indicating more wrong vessels are captured).

From Table 3, we observe that the overall performance is best when the spacing is close to the original spacing (0.6mm), and spacing 0.3 mm gives better results than 0.9mm.

5 Conclusion

In this paper, we combined attention mechanism to improve the 3D U-Net model. Also, the impact of different loss and isotropic spacing on the performance of coronary artery tree segmentation was investigated. It can be seen that among the improved models, 3D U-Net with CAM has the best segmentation performance, especially it yields highest sensitivity but still keeps segmentation clean. From the experiments of focal loss, we should pay more attention to the problem of unbalanced segmentation during training, specifically, in the coronary artery tree segmentation, loss of the foreground class to participate in the gradient calculation. Regarding resampling spacing, isotropic spacing similar to the original spacing for training gives the best results.

References

- [1]. Zhao, C., Tang, H., Tang, J., Zhang, C., He, Z., Wang, Y. P., ... & Zhou, W. (2020). Semantic Segmentation to Extract Coronary Arteries in Fluoroscopy Angiograms. medRxiv.
- [2] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
- [3] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Springer, Cham.
- [4]. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3146-3154).
- [5]. Merkow, J., Marsden, A., Kriegman, D., & Tu, Z. (2016, October). Dense volume-to-volume vascular boundary detection. In International conference on medical image computing and computer-assisted intervention (pp. 371-379). Springer, Cham.
- [6]. Moeskops, P., Wolterink, J. M., van der Velden, B. H., Gilhuijs, K. G., Leiner, T., Viergever, M. A., & Išgum, I. (2016, October). Deep learning for multi-task medical image segmentation in multiple modalities. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 478-486). Springer, Cham.

- [7]. Shen, Y., Fang, Z., Gao, Y., Xiong, N., Zhong, C., & Tang, X. (2019). Coronary arteries segmentation based on 3D FCN with attention gate and level set function. *IEEE Access*, 7, 42826-42835.
- [8]. Xu, C., Xu, L., Gao, Z., Zhao, S., Zhang, H., Zhang, Y., ... & Li, S. (2018). Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. *Medical image analysis*, 50, 82-94.
- [9]. Lee, M. C. H., Petersen, K., Pawlowski, N., Glocker, B., & Schaap, M. (2019). TETRIS: template transformer networks for image segmentation with shape priors. *IEEE transactions on medical imaging*, 38(11), 2596-2606.
- [10]. Chen, Y. C., Lin, Y. C., Wang, C. P., Lee, C. Y., Lee, W. J., Wang, T. D., & Chen, C. M. (2019). Coronary artery segmentation in cardiac CT angiography using 3D multi-channel U-Net. *arXiv preprint arXiv:1907.12246*.
- [11]. Duan, Y., Feng, J., Lu, J., & Zhou, J. (2018, September). Context aware 3D fully convolutional networks for coronary artery segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart* (pp. 85-93). Springer, Cham.
- [12]. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Glocker, B. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.
- [13]. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057).
- [14]. Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4995-5004).
- [15]. Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21-29).
- [16]. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5659-5667).
- [17]. Tolooshams, B., Giri, R., Song, A. H., Isik, U., & Krishnaswamy, A. (2020, May). Channel-Attention Dense U-Net for Multichannel Speech Enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 836-840). IEEE.
- [18]. Chen, Y. C., Lin, Y. C., Wang, C. P., Lee, C. Y., Lee, W. J., Wang, T. D., & Chen, C. M. (2019). Coronary artery segmentation in cardiac CT angiography using 3D multi-channel U-net. *arXiv preprint arXiv:1907.12246*.
- [19]. Livne, M., Rieger, J., Aydin, O. U., Taha, A. A., Akay, E. M., Kossen, T., ... & Madai, V. I. (2019). A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Front Neurosci* 13 (97).
- [20]. Kong, B., Wang, X., Bai, J., Lu, Y., Gao, F., Cao, K., ... & Yin, Y. (2020). Learning tree-structured representation for 3D coronary artery segmentation. *Computerized Medical Imaging and Graphics*, 80, 101688.
- [21]. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

- [22]. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.
- [23]. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).
- [24]. Peng, C., Zhang, X., Yu, G., Luo, G., & Sun, J. (2017). Large kernel matters--improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4353-4361).