

IarCAC: Instance-aware Representation for Coronary Artery Calcification Segmentation in Cardiac CT angiography

Weili Jiang¹, Yiming Li², Zhang Yi¹, Jianyong Wang¹ , and Mao Chen² 

¹ Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, Sichuan, P. R. China

² Department of Cardiology, West China Hospital, Sichuan University, Chengdu, Sichuan, P. R. China

 *corresponding authors* wjy@scu.edu.cn (Jianyong Wang),
hmaochen@vip.sina.com (Mao Chen)

Abstract. Coronary Artery Calcification (CAC) is a robust indicator of coronary artery disease and a critical determinant of percutaneous coronary intervention outcomes. Our method is inspired by a clinical observation that CAC typically manifests as a sparse distribution of multiple instances. Existing methods focusing solely on spatial correlation overlook the sparse spatial distribution of semantic connections in CAC tasks. Motivated by this, we introduce a novel instance-aware representation method for CAC segmentation, termed IarCAC, which explicitly leverages the sparse connectivity pattern among instances to enhance the model’s instance discrimination capability. The proposed IarCAC first develops an InstanceViT module, which assesses the connection strength between each pair of tokens, enabling the model to learn instance-specific attention patterns. Subsequently, an instance-aware guided module is introduced to learn sparse high-resolution representations over instance-dependent regions in the Fourier domain. To evaluate the effectiveness of the proposed method, we conducted experiments on two challenging CAC datasets and achieved state-of-the-art performance across all datasets. The code is available at <https://github.com/WeiliJiang/IarCAC>.

Keywords: Coronary Artery Calcification Segmentation · InstanceViT · Instance-aware Guided.

1 Introduction

Cardiovascular disease (CVD) is one of the leading causes of mortality worldwide [11]. Coronary artery calcification (CAC) has been identified as a significant independent predictor of cardiovascular events [14]. Therefore, accurate segmentation of CAC holds significant importance for the prediction of CVD.

Clinically, CAC is quantified through coronary artery calcium scoring computed tomography (CSCT). Previous studies have introduced many traditional methods in CSCT for CAC segmentation, such as Nearest Neighbors [9], Support

Vector Machines [24], Random Decision Trees [20], and [23,1] focus on coronary artery segmentation and then combine coronary artery information and voxel intensity values to identify CAC. However, studies [16] have demonstrated the potential use of cardiac CT angiography (CCTA) for CAC quantification, and Using CCTA for CAC analysis has the potential to reduce the radiation dose associated with cardiac CT exams by approximately 40-50% [22]. Traditional methods cannot be applied in CCTA, as they classify potential CAC lesions extracted using a clinical 130 HU threshold. In CCTA, it is non-trivial to distinguish between CAC and attenuated lumen, and the application of a predefined single detection threshold to extract potential CAC lesions is not feasible. Instead, with the advent of deep learning models and the innovative design of network architectures, these techniques segment lesions by identifying CAC voxels. These voxel classification methods primarily revolve around two categories: Convolutional Neural Networks (CNNs) [19,25] and Transformers [21]. Specifically, CNNs encompass architectures like Unet [19,15] and its derivatives, including the 3D pyramid pooling network [28], ResNet-3D [7], spatial-temporal encoder-decoder [13], and nnUnet [8]. On the other hand, Transformers employ the global self-attention mechanism to yield variants like U-Transformer [17], Swin Transformer [6], and nnFormer [27].

While the aforementioned methods have shown promising results, they still face limitations, making CAC segmentation a challenging task due to its variable sizes and shapes, low contrast, and high noise characteristics. On the one hand, CNN-based approaches [19,13,7] struggle with capturing long-range dependencies due to their limited local receptive fields, thus compromising segmentation accuracy. On the other hand, methods that combine CNNs and Transformers [17,6,27] aim to address this issue but still face representational limitations and increased noise tokens due to the fixed attention pattern of full self-attention computation. Given the sparse spatial distribution of CAC within the image, it raises the question: *Is it necessary to represent all content for effective CAC segmentation?*

During segmentation, redundant image tokens, especially those containing only background information, often fail to contribute meaningful contextual data. Therefore, concentrating solely on instance-related regions is sufficient for precise estimation. Drawing from this insight, we introduce IarCAC, an **I**nstance **a**ware **R**epresentation framework designed for CAC segmentation. This method harmoniously integrates CNNs and Transformers. Our approach is guided by two principles: (1) Establishing instance-aware sparse patterns. Unlike conventional self-attention patterns, these sparse patterns encourage tokens to utilize their limited non-zero attention budget more effectively. In semantic-focused attention heads, tokens carrying similar semantic content should exhibit higher connectivity scores, regardless of spatial proximity. (2) Developing an instance-aware guided matching module. Operating in the Fourier domain, this module discriminates which low- and high-frequency information around instances should be retained for precise segmentation outcomes.

Overall, the following are the contributions of this work:

- We propose an instanceViT to capture the variable distribution of semantic information within instances in the input image content.
- We introduce an instance-aware guided module for learning sparse high-resolution representations over instance-dependent regions.
- Extensive experiments are performed on two CAC datasets, resulting in new state-of-the-art performances consistently.

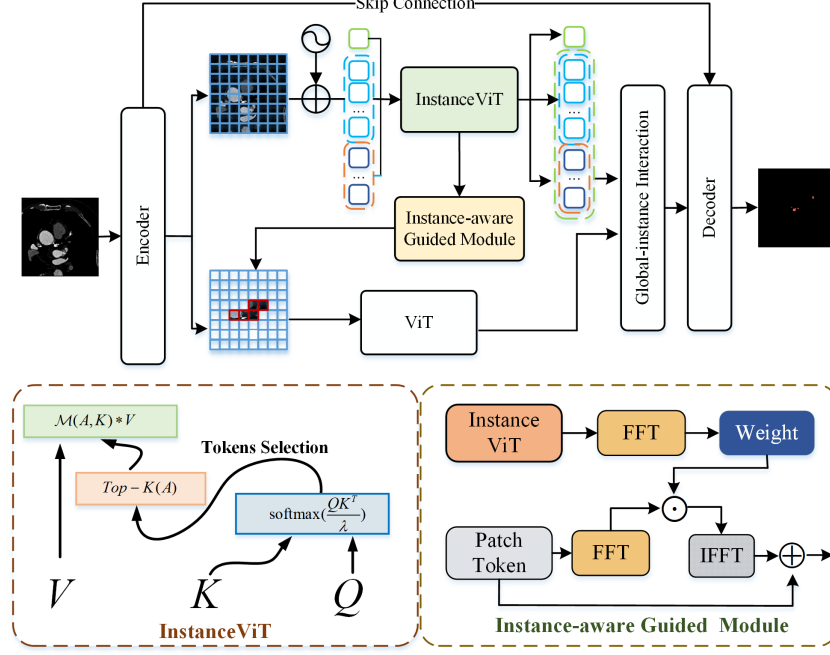


Fig. 1. The overall structure of IarCAC.

2 Methods

The overall architecture of IarCAC is depicted in Fig. 1, mainly comprising encoders, instance-aware learning, and decoders. In this section, we will present our framework stage-by-stage and give a detailed introduction to each module.

Encoder and Decoder. The encoder of IarCAC comprises four encoding modules, with each module containing two series of "convolution-normalization-activation" blocks. A MaxPooling layer is employed between the modules for downsampling the feature maps and enlarging the receptive field. The decoder, likewise, consists of four decoder modules, with each module incorporating two series of "convolution-normalization-activation" blocks and an upsampling layer. This upsampling layer elevates the resolution of the input channels by utilizing

bilinear interpolation. Furthermore, inspired by U-Net [19], we introduce symmetric skip connections between corresponding feature pyramids of the encoder and decoder, facilitating the recovery of fine-grained details in the predictions.

Instance-aware learning. Instance-aware learning encompasses two distinct stages. Initially, an instanceViT is formulated to dynamically assess the interplay between image regions and instances, generating preliminary estimates of instances. Subsequently, an instance-aware guided module is employed to assess the necessity for refining these preliminary estimates. To bolster instance-specific dependencies, we introduce a global-instance interaction module. This module integrates instance-specific features with global features through concatenation, followed by a block consisting of two 3×3 convolutional layers, batch normalization, ReLU, and a sigmoid layer.

2.1 InstanceViT

Our method is based on the existing ViT architecture and a simple implementation of Multi-Head Self-Attention (MHSA). Since the standard ViT [21] relies on all tokens for computing global self-attention, it is not conducive to CAC segmentation as it may lead to undesirable interactions between unrelated features, introducing noise. To address these limitations, we leverage the sparse distribution of instances within the image space and develop InstanceViT as a feature extraction component. Formally, given the input features at $(l - 1)$ -th block X_{l-1} , the encoding process of InstanceViT can be defined as follows:

$$\begin{cases} X'_l = X_{l-1} + \text{IASA}(\text{LN}(X_{l-1})) \\ X_l = X'_l + \text{FFN}(\text{LN}(X'_l)) \end{cases}, \quad (1)$$

where LN denotes the layer normalization; X'_l and X_l denote the outputs from instance-aware sparse attention (IASA) and feed-forward network (FFN). The FFN applies a fully connected network to each position in the sequence, adding non-linearity and enhancing the model's ability to capture complex patterns.

IASA generates connectivity attention scores as $A(Q, K) = \frac{QK^T}{\sqrt{d}}$, where the query matrix is $Q = X_i^l K^Q$ and the key matrix is $K = X_i^l W^K$, the projection matrices W^Q and W^K are learnable parameters projecting the input feature x^l at layer l . To implement IASA with limited connections, we utilize a sparse attention masking operation denoted as $\mathcal{M}(\cdot)$ on the attention score matrix A , selecting the top- k contributing elements. Specifically, we identify the k -th largest element in each row of A and record the positions (i, j) in the location matrix, where k is a user-defined hyper-parameter. Assuming the k -th threshold value in the i -th row is t_i , the position (i, j) is recorded if the value of the j -th component surpasses t_i . Concatenating the thresholds of each row forms a vector $t = [t_1, t_2, \dots, t_{l_Q}]$. The sparse connectivity mask function $\mathcal{M}(\cdot, \cdot)$ is derived as follows:

$$\mathcal{M}(A, k)_{ij} = \begin{cases} A_{ij}, & \text{if } A_{ij} \geq t_i (k\text{-th largest value of row } i) \\ 0, & \text{if } A_{ij} < t_i (k\text{-th largest value of row } i) \end{cases}. \quad (2)$$

This dynamic instance-aware sparse tokens selection from *dense* to *instance*, as demonstrated by:

$$\text{InstanceAtt} = \text{softmax}(\mathcal{M}(A, K))V. \quad (3)$$

2.2 Instance-aware Guided Module

Based on the spectral convolution theorem in Fourier theory, it is noted that pointwise updates in the Fourier domain possess the capability to globally influence all input features, as discussed in prior research by [3]. This observation underscores how spectral learning facilitates the capture of instance-based global interactions across all frequencies [18]. Motivated by this insight, we delve into the exploration of instance-aware guided interactions within the Fourier domain, departing from previous endeavors primarily centered on spatial domain interactions [2,5]. The architecture of the Instance-aware Guided Module is presented in Fig. 1, with detailed explanations of each step provided below.

The instance-aware interaction dynamically identifies the frequency information pertinent to the instance region, allowing another ViT to assess the necessity of refining the initial estimations. Specifically, the spatial features of the instanceViT, denoted as F_{ins} , are transformed into the Fourier domain. As proved in [18], multiplying the spectrum with global weights can effectively exchange spatial information. Adhering to this approach, we learn interaction weights $\mathcal{K}_{ins}$ from the instanceViT branch to modulate the representation of ViT. To generate these interaction weights, we adopt a simple block of resnet and a sigmoid layer. Following the computational efficiency practices outlined in [18], we calculate the weights solely using the real part of the complex spectrum. Subsequently, we leverage the interaction weights $\mathcal{K}_{ins}$ to guide F_{ViT} in Fourier domain. The $\mathcal{K}_{ins}$ is obtained by a series of two 1×1 convolution layers, BN and GELU. Then, the Inverse Fast Fourier Transform (IFFT) ($\mathcal{F}^{-1}$) converts the guided features back into the spatial domain $\tilde{F}_{ViT}$ and combines with F_{ViT} to form a residual path:

$$\hat{F}_{ViT} = \mathcal{F}^{-1} [\mathcal{K}_{ins} (\mathcal{F} (F_{ins})) \odot \mathcal{F} (F_{ViT})] \oplus F_{ViT}. \quad (4)$$

2.3 Loss Function

To accomplish the segmentation task, we employ a complimentary combination of the widely utilized soft dice loss and cross-entropy loss. This integrated approach allows us to harness the strengths of both loss functions. It is defined as:

$$\mathcal{L}(Y, P) = 1 - \sum_{i=1}^I \left(\frac{2 \sum_{v=1}^V Y_{v,i} \cdot P_{v,i}}{\sum_{v=1}^V Y_{v,i}^2 + \sum_{v=1}^V P_{v,i}^2} + \sum_{v=1}^V Y_{v,i} \log P_{v,i} \right), \quad (5)$$

where I is the number of classes; V is the number of voxels; $Y_{v,i}$ and $P_{v,i}$ denote the ground truths and output probabilities at voxel v for class i , respectively.

3 Experiments

3.1 Datasets and Data-processing

To verify the efficacy of our model, we select the two CAC datasets. Both datasets were independently and anonymously annotated by cardiologists and radiologists using 3D Slicer software. Dice between each annotation and its union measured the annotator’s preference. Inconsistencies were rechecked. For training and evaluation purposes, the dataset is divided into two subsets. Specifically, 80% of the images are allocated for training, while 20% of the images are set aside for testing.

CAC-CTA dataset. The Chinese top-grade hospital collected the CTA data of 150 patients [10]. The dataset contains 802 individual instances, averaging 5.36 ± 4.92 instances per sample.

ImageCAS-CAC dataset. The dataset includes the first 150 patients from the ImageCAS dataset [26]. The dataset contains 831 individual instances, averaging 5.54 ± 8.22 instances per sample.

Data-processing of CAC dataset. The CTA images are interpolated to the same thickness (i.e., 0.5mm). Furthermore, we used a threshold range from -224HU to 600HU for coarse-segment the lung, then used the seed-filling algorithm to fine-segment the lung. Subtracting the lung from the original image can eliminate noise to better segment CAC.

Table 1. Quantitative comparison on CAC-CTA and ImageCAS-CAC datasets. The best results are boldfaced, and the second-best results are underlined.

Dataset	method	IF1	IS	IP	DSC	SDSC
CAC-CTA	Unet [4]	0.689 $\pm$ 0.10	<u>0.741$\pm$0.16</u>	0.692 $\pm$ 0.16	0.689 $\pm$ 0.10	0.535 $\pm$ 0.11
	U-Transformer [17]	0.632 $\pm$ 0.11	0.654 $\pm$ 0.20	0.703 $\pm$ 0.19	0.661 $\pm$ 0.15	0.489 $\pm$ 0.16
	SwinUNTER [6]	0.687 $\pm$ 0.11	0.703 $\pm$ 0.17	0.702 $\pm$ 0.16	0.676 $\pm$ 0.11	0.522 $\pm$ 0.12
	nnFormer [27]	0.643 $\pm$ 0.18	0.735 $\pm$ 0.17	0.653 $\pm$ 0.20	0.656 $\pm$ 0.16	0.497 $\pm$ 0.17
	nnUnet [8]	0.700 $\pm$ 0.09	0.744$\pm$0.17	<u>0.706$\pm$0.14</u>	<u>0.691$\pm$0.09</u>	0.548 $\pm$ 0.11
	IarCAC(Ours)	0.713$\pm$0.11	0.731 $\pm$ 0.16	0.723$\pm$0.13	0.723$\pm$0.11	0.584$\pm$0.09
ImageCAS-CAC	Unet [4]	0.783 $\pm$ 0.11	0.677 $\pm$ 0.12	0.769 $\pm$ 0.11	0.673 $\pm$ 0.12	0.548 $\pm$ 0.15
	U-Transformer [17]	<u>0.796$\pm$0.15</u>	0.623 $\pm$ 0.13	0.806 $\pm$ 0.17	0.681 $\pm$ 0.18	0.549 $\pm$ 0.13
	SwinUNTER [6]	0.806$\pm$0.13	0.683 $\pm$ 0.12	0.802 $\pm$ 0.15	0.697 $\pm$ 0.21	<u>0.561$\pm$0.08</u>
	nnFormer [27]	0.741 $\pm$ 0.14	0.604 $\pm$ 0.18	0.759 $\pm$ 0.15	0.654 $\pm$ 0.17	0.486 $\pm$ 0.13
	nnUnet [8]	0.769 $\pm$ 0.15	<u>0.721$\pm$0.13</u>	<u>0.806$\pm$0.19</u>	<u>0.703$\pm$0.11</u>	0.553 $\pm$ 0.13
	IarCAC(Ours)	0.789 $\pm$ 0.12	0.728$\pm$0.15	0.811$\pm$0.17	0.717$\pm$0.13	0.601$\pm$0.12

3.2 Implementation Details and Evaluation Metrics

• **Implementation Details.** The models were trained on an NVIDIA GeForce RTX 3090 with the Pytorch framework, employing the SGD optimizer. The initial learning rate was set to 0.001, with a weight decay of $2e-4$. We utilized the ReduceLROnPlateau mechanism with a coefficient of 0.5, a patience and cooldown of 3, and a minimum learning rate of $1e-8$. For data preprocessing, a

$192 \times 192 \times 192$ patch was randomly cropped from each CCTA image, normalized, and fed into the neural network. During training, a batch size of 2 was used, with 12 worker threads, and each experiment ran for 120 epochs.

• **Evaluation Metric.** To assess the overall volumetric performance, we employed the Dice Similarity Coefficient (DSC) and the Surface Dice Similarity Coefficient (SDSC). To further validate the evaluation metrics specifically for CAC, we adopted instance-wise detection metrics, namely Instance F1 (IF1), Instance Sensitivity (IS), and Instance Precision (IP), following the methodology of prior work [12]. In addition, the segmentation threshold is 0.5.

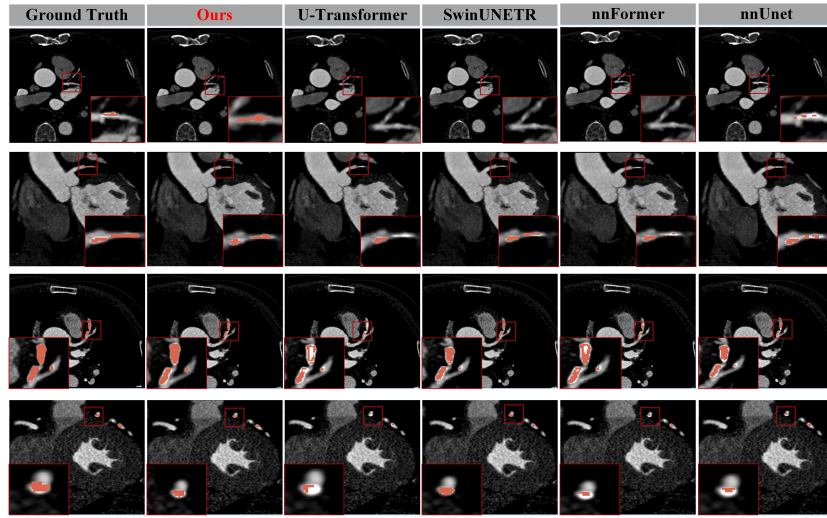


Fig. 2. Visual comparison of compared methods on CAC-CTA and ImageCAS-CAC datasets.

3.3 Comparison with State-of-the-Art Methods

We first compare the proposed model with five state-of-the-art (SOTA) methods: Unet [4], U-Transformer [17], SwinUNETR [6], nnFormer [27], and nnUnet [8]. The evaluation results on the CAC-CTA and ImageCAS-CAC datasets are presented in Table 1. Compared to U-Transformer, SwinUNETR demonstrates significant performance gains through a multi-level feature fusion mechanism, which illustrates the effectiveness of feature fusion in the context of medical image segmentation. In addition, nnUnet stands out for its adaptability, offering flexible structural and hyper-parameter adjustments tailored to specific task requirements. However, for the task of CAC segmentation, the integration of clinical prior knowledge is crucial to enhance the instance detection capability. Our method exploits the sparse connection pattern of instances to enhance

the model’s instance detection capabilities, and in contrast to these methods, our proposed model exhibits improvements in both global volumetric indicators (DSC, SDSC) and instance-level indicators (IF1, IS, IP).

Fig. 2 presents a visual comparison of segmentation results obtained by various methods. Our proposed approach stands out for its superior accuracy, producing smoother and sharper edges compared to other methods. Notably, the compared methods are prone to exhibit discrete mispredictions and missed the detection of finer details, particularly in the case of smaller objects.

Table 2. The impact of Hyper-parameter top- K in InstanceViT.

Top-K	IF1	IS	IP	DSC	SDSC
4	0.673±0.17	0.696±0.20	0.704±0.19	0.673±0.17	0.520±0.16
8	0.713±0.11	0.731±0.16	0.723±0.13	0.723±0.11	0.584±0.09
12	0.713±0.13	0.754±0.15	0.705±0.16	0.712±0.13	0.581±0.13
16	0.710±0.12	0.746±0.16	0.717±0.12	0.710±0.12	0.560±0.11
20	0.699±0.11	0.691±0.17	0.761±0.15	0.699±0.11	0.521±0.11
24	0.699±0.13	0.732±0.15	0.713±0.16	0.699±0.13	0.525±0.10

3.4 Ablation Study

• **The impact of Hyper-parameter top- k .** To investigate the influence of varying k within the InstanceViT, we conducted an ablation study. The range of top- k values considered spans from 4 to 24, as detailed in Table 2. Notably, when top- k is set to 8, the model demonstrates its peak segmentation performance.

Table 3. Ablation study of our model on CAC-CTA dataset. InsViT denotes the InstanceViT module, and Guided denotes the instance-aware guided module.

Unet	InsViT	Guided	IF1	IS	IP	DSC	SDSC
✓			0.689±0.10	0.741±0.16	0.692±0.16	0.689±0.10	0.535±0.11
✓	✓		0.702±0.11	0.750±0.16	0.716±0.15	0.716±0.11	0.553±0.15
✓		✓	0.709±0.15	0.738±0.16	0.719±0.18	0.715±0.15	0.565±0.16
✓	✓	✓	0.713±0.11	0.731±0.16	0.723±0.13	0.723±0.11	0.584±0.09

• **Effectiveness of Individual Components of the proposed IarCAC**
To evaluate the effectiveness of each component within the proposed model, we performed an ablation study on the CTA-CAC dataset. First, the InstanceViT module was utilized to find instance-related tags. Then, the Instance-aware Guided matching module was employed to direct the global semantic information to high-resolution representation. Table 3 demonstrate that each component significantly impacts the overall performance, highlighting the importance of both the InstanceViT and Instance-aware Guide modules.

4 Conclusion

In this paper, we introduce a novel model tailored for the CAC segmentation challenge. By acknowledging the sparsity of CAC in the image domain, we introduce the InstanceViT to assess the connectivity score between each pair of tokens, facilitating the learning of instance-specific attention patterns. Furthermore, we investigate instance-aware guided global semantic learning in the Fourier domain, leveraging spectral learning’s ability to effectively direct frequencies towards capturing long-term interactions. Our proposed model exhibits impressive performance on two challenging CAC tasks. In summary, our method enables accurate and automated CAC identification and quantification within CCTA. This advancement may eliminate the need for a dedicated CSCT scans, traditionally acquired prior to CCTA, thereby reducing the radiation dose received by patients.

Acknowledgement. This work was supported by the National Major Science and Technology Projects (Grant No. 2018AAA0100201); the National Natural Science Foundation of China (Grant No. 81970325 and Grant No. 62306192); Post doctor fellow support fund from Sichuan University (Grant No. 20826041E4070); Open Research Fund Program of Data Recovery Key Laboratory of Sichuan Province (Grant No. DRN2201); Natural Science Foundation of Sichuan Province (Grant No. 2023NSFSC1638).

Disclosure of Interests. The authors have no competing interests.

References

1. Ahmed, W., de Graaf, M.A., Broersen, A., Kitslaar, P.H., Oost, E., Dijkstra, J., Bax, J.J., Reiber, J.H., Scholte, A.J.: Automatic detection and quantification of the agatston coronary artery calcium score on contrast computed tomography angiography. *The international journal of cardiovascular imaging* **31**, 151–161 (2015)
2. Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., Liu, Z.: Mobile-former: Bridging mobilenet and transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5270–5279 (2022)
3. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. *Advances in Neural Information Processing Systems* **33**, 4479–4488 (2020)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II* 19. pp. 424–432. Springer (2016)
5. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12175–12185 (2022)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI Brainlesion Workshop*. pp. 272–284. Springer (2021)

7. He, C., Wang, J., Yin, Y., Li, Z.: Automated classification of coronary plaque calcification in oct pullbacks with 3d deep neural networks. *Journal of Biomedical Optics* **25**(9), 095003 (2020)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
9. Isgum, I., Prokop, M., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Automatic coronary calcium scoring in low-dose chest computed tomography. *IEEE Transactions on Medical Imaging* **31**(12), 2322–2334 (2012)
10. Jiang, W., Li, Y., Jia, Y., Feng, Y., Yi, Z., Chen, M., Wang, J.: Ori-net: Orientation-guided neural network for automated coronary arteries segmentation. *Expert Systems with Applications* **238**, 121905 (2024)
11. Kochanek, K.D., Murphy, S.L., Xu, J., Arias, E.: Deaths: final data for 2017 (2019)
12. Kofler, F., Shit, S., Ezhov, I., Fidon, L., Horvath, I., Al-Maskari, R., Li, H.B., Bhatia, H., Loehr, T., Piraud, M., et al.: blob loss: instance imbalance aware loss functions for semantic segmentation. In: *International Conference on Information Processing in Medical Imaging*. pp. 755–767. Springer (2023)
13. Li, C., Jia, H., Tian, J., He, C., Lu, F., Li, K., Gong, Y., Hu, S., Yu, B., Wang, Z.: Comprehensive assessment of coronary calcification in intravascular oct using a spatial-temporal encoder-decoder network. *IEEE Transactions on Medical Imaging* **41**(4), 857–868 (2021)
14. Moussa, I.D., Mohanane, D., Saucedo, J., Stone, G.W., Yeh, R.W., Kennedy, K.F., Waksman, R., Teirstein, P., Moses, J.W., Simonton, C.: Trends and outcomes of restenosis after coronary stent implantation in the united states. *Journal of the American College of Cardiology* **76**(13), 1521–1531 (2020)
15. Mu, D., Bai, J., Chen, W., Yu, H., Liang, J., Yin, K., Li, H., Qing, Z., He, K., Yang, H.Y., et al.: Calcium scoring at coronary ct angiography using deep learning. *Radiology* **302**(2), 309–316 (2022)
16. Pavitt, C.W., Harron, K., Lindsay, A.C., Ray, R., Zielke, S., Gordon, D., Rubens, M.B., Padley, S.P., Nicol, E.D.: Deriving coronary artery calcium scores from ct coronary angiography: a proposed algorithm for evaluating stable chest pain. *The international journal of cardiovascular imaging* **30**, 1135–1143 (2014)
17. Petit, O., Thome, N., Rambour, C., Themyl, L., Collins, T., Soler, L.: U-net transformer: Self and cross attention for medical image segmentation. In: *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings* 12. pp. 267–276. Springer (2021)
18. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. *Advances in neural information processing systems* **34**, 980–993 (2021)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. pp. 234–241. Springer (2015)
20. Shahzad, R., van Walsum, T., Schaap, M., Rossi, A., Klein, S., Weustink, A.C., de Feyter, P.J., van Vliet, L.J., Niessen, W.J.: Vessel specific coronary artery calcium scoring: an automatic system. *Academic Radiology* **20**(1), 1–9 (2013)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
22. Voros, S., Qian, Z.: Agatston score tried and true: by contrast, can we quantify calcium on cta? *Journal of Cardiovascular Computed Tomography* **6**(1), 45–47 (2012)

23. Wang, W., Yang, L., Wang, S., Wang, Q., Xu, L.: An automated quantification method for the agatston coronary artery calcium score on coronary computed tomography angiography. *Quantitative Imaging in Medicine and Surgery* **12**(3), 1787 (2022)
24. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 589–596. Springer (2015)
25. Wolterink, J.M., Leiner, T., de Vos, B.D., van Hamersvelt, R.W., Viergever, M.A., Išgum, I.: Automatic coronary artery calcium scoring in cardiac ct angiography using paired convolutional neural networks. *Medical image analysis* **34**, 123–136 (2016)
26. Zeng, A., Wu, C., Huang, M., Zhuang, J., Bi, S., Pan, D., Ullah, N., Khan, K.N., Wang, T., Shi, Y., et al.: Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *arXiv preprint arXiv:2211.01607* (2022)
27. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24. pp. 14–24. Springer (2021)
28. Zhou, R., Guo, F., Azarpazhooh, M.R., Spence, J.D., Ukwatta, E., Ding, M., Fenster, A.: A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3d ultrasound images. *IEEE Transactions on Medical Imaging* **39**(9), 2844–2855 (2020)