Real-time Segmentation of Coronary Artery Calcification Using Spatial Attention and Parallel Convolution

Tetsuya Asakawa* Yuki Sugimoto** Masashi Hashimoto‡ Takeshi Miyaji§ Kazuki Shimizu¶ Kei Nomura‖ and Masaki Aon
*†** Toyohashi University of Technology, Aichi, Japan
‡§¶‖ Toyohashi Heart center, Aichi, Japan
* E-mail: asakawa@kde.cs.tut.ac.jp
** E-mail: sugimoto.yuki.fd@tut.jp
‡ E-mail: m.hashimo6123@gmail.com
§ E-mail: miyaji@heart-center.or.jp
¶ E-mail: K
‖ E-mail: k.nomura@heart-center.or.jp
† E-mail: masaki.aono.ss@tut.ac.jp

***Abstract*—Semantic segmentation is an image recognition technique that classifies each pixel in an image. This technique is crucial in the fields of automated driving and medicine. Despite its significance, limited computational resources is available for its application. The current challenge involves to developing models that are accurate and fast, secondary lightweight. DDRNet (Deep Dual-Resolution Net) is a real-time segmentation method, which is a highly accurate method with a multi-resolution substructure. DDRNet is a model that to achieve efficiently both accuracy and real-time performance reducing the number of resolutions used in the subnetworks. However, DDRNet is less accurate than non-lightweight models, especially in small object recognition. Therefore, this research aimed to improve the accuracy of DDRNet while maintaining its light-weight and high computation speed. We conducted experiments on a coronary calcification dataset using DDR-SAPC-U as our proposed model, which is a modified version of the proposed method specialized for medical imaging with the addition of a skip connection. Overall, DDR-SAPC-U as our proposed demonstrated the highest accuracy compared to conventional medical methods and was relatively lightweight in terms of computational complexity. DDR-SAPC-U as our proposed model improved the accuracy compared to conventional methods.**

I. Introduction

Semantic segmentation is an image recognition technique that classifies each pixel in an image. This technology is widely applied role in various fields. Such as object recognition in autonomous driving, abnormal part detection in industrial inspection, and detection of organs and lesions in the medical field. In cardiac CT, this technology is applied for detailed detection of lesions called coronary artery calcification, which are strongly associated with coronary artery disease. However, the limited computational resources in cars and medical settings have prompted advancements in semantic segmentation using deep learning wherein the development of a model focuses not only on accuracy but also on lightness and speed.

HRNet (High-Resolution Net) [1] [2] is a powerful backbone in image recognition tasks, and methods applying it have recorded high accuracy in semantic segmentation, pose estimation, and object detection. Owing to the structure in which the sub-networks with multiple resolutions are connected in parallel, the computation amount is large and the processing speed is slow. DDRNet [3] is a real-time segmentation method inspired by HRNet and has a structure in which sub-networks with two different resolutions are connected in parallel. DDRNet is a model that to achieve efficiently both accuracy and real-time performance reducing the number of resolutions used in the subnetworks. However, DDRNet is less accurate than non-lightweight models, especially in small object recognition, especially in images. Therefore, this research aims to improve the accuracy of DDRNet by using coronary artery data while maintaining its lightweight nature and high computation speed.

The remainder of this paper is organized as follows. In Section 2, as related to research, we will explain a semantic segmentation method based on deep learning, a method using attention, one using multiple resolutions, and another for medical images. The describes the proposed method is described in Section 3. Results of comparative experiments between the proposed method and the conventional method are discussed in Section 4. In Section 5, we discuss the details of the experiment and conclusions and suggestions for further work are summarized in Section 6.

II. Related Work

In this Section, we review research literature related to the topic of this study. First, we describe research on medical images in semantic segmentation using deep learning. Thereafter, research on methods using attention and methods using multiple resolutions are described. Finally, we describe DDRNet, which is the baseline of this research.

A. Semantic segmentation

Semantic segmentation uses deep learning developed from FCN [4]. Thus far, DeepLab [5], [6], which improved accuracy by convolution using U-Net with Encoder-Decoder structure and multiple dilated rates. In addition, several methods have been proposed, such as PSPNet [7], which improves accuracy by using multiple sizes of Average Pooling.

Currently, numerous segmentation methods that pursue real-time performance. For example, SegNet [8], which is accelerated by using a small network structure and skip coupling, and spatial information is extracted from two networks called Spatial Path and Context Path, BiSeNet [9], and BiSeNet v2 [10], etc. The other studies, E-Net [11], ICNet [12], and SFNet [13] have proposed fast and lightweight semantic segmentation methods.

B. Medical images in semantic segmentation

Semantic segmentation has a great impact in the medical field, and is useful for detecting organs and lesions. UNet [14] is a simple Encoder-Decoder model featuring skip joins. FCN, the main semantic segmentation method proposed before UNet, yielded only low-resolution output.

On the other hand, UNet has an Encoder-Decoder structure, and an output with the same resolution as the input can be obtained. For this reason, it is often used in tasks in the medical field that require dense output, and has recently been applied to the detection of the brain, lungs, and heart. Some models based on the structure of UNet have also been proposed. Examples include UNet++ [15], which has a dense structure with multiple convolutions and skip connections added between the encoder and decoder, and TransUNet [16], which incorporates a transformer into the structure of UNet. These encoder-decoder structure methods have recorded high accuracy in tasks in the medical field.

C. Attention mechanism

In recent years, attention mechanisms [17] have been introduced not only in natural language processing but also in semantic segmentation. DANet [18] and PSA [19] are methods that focus on this. These methods have a network structure that uses both spatial and channel-oriented attention. In particular, PSA recorded high accuracy by combining OCR [20], which is a method that also applies attention, and HRNet, which has multiple resolution sub-networks.

In addition, Tao et al. [21] proposed a method that outputs spatial direction attention at multiple resolutions, and by using spatial attention, it was possible to extract information more appropriately for object regions than conventional convolutions. In this method, the element product of each attention map and feature map is taken to emphasize the area of the object and improve the accuracy. From this research, we obtained findings showing that the object to be focused on differs depending on the resolution of the image used.

III. Proposed Method

We propose two methods, Dual Attention Module (DAM) and Parallel Basic Block (PBB), which effectively extract features with less computation. Furthermore, we propose ding DDR-Spatial Attention Parallel Convolution (DDR-SAPC) to DDRNet-23-slim.

A. DDR-SAPC-U

We improved DDR-SAPC for application to medical image datasets. The improved network is called DDR-SAPC-U. Figure 1 is a schematic of the network of DDR-SAPC-U. There are two improvements from DDR-SAPC. The first point is the resolution of convolution. Even a small error in medical image segmentation can greatly change the diagnostic result. In addition, detection of lesions and abnormalities requires high accuracy. Therefore, we modified the Encoder-Decoder structure to output at the original resolution such as U-Net [14] and UNet++ [15]. This structure enables the fine detection of objects, and further improvement in accuracy can be expected. The second point is the addition of skip joins. In the encoder-decoder structure, the loss of features obtained on the encoder side occurs during convolution and upsampling. To prevent this, a skip join is introduced in the red-lined part of the figure. The introduction of skip coupling prevents the loss of detailed features. Furthermore, the detection performance is expected to improve.

1) Dual Attention Module: The Dual Attention Module (DAM) is a module inspired by Tao's method [21]. Figure 2 displays the network structure.

This module is decomposed into the part that generates spatial attention in Figure 2(a) and the part that extracts features by convolving feature maps in Figure 2(b). In DDR-Net, fusion of low and high-resolution was performed only in the part displayed in Figure 2(b). Therefore, we added a spatial attention generation part emphasizing the feature values of objects obtained at low and high-resolution. In the spatial attention generator, low-resolution and high-resolution attention maps AS and AL are generated by two layers of 3×3 conv, respectively. Then, $\mathrm{A_H}$+(1-A_L) for each element on the high-resolution side and $\mathrm{A_L} + (1 - \mathrm{A_H})$ on the low-resolution side is processed. The element product considers the resulting attention map and the original feature map. Generate a feature map by passing it through two layers of convolution and adding it to the inverse resolution.

2) Parallel Basic Block: PBB is a module devised to increase the types of resolutions in the convolutional layers inside DDRNet. DDRNet-23-slim uses the Basic Block in Figure 2(a) for the internal convolution block. This method is used in ResNet18 [22] etc. It is a very lightweight block that extracts features with two layers of convolution.

In contrast, the proposed method, Parallel Basic Block (PBB), has a structure in which two layers of convolution are performed for three kinds of resolutions [1/1, 1/2, 1/4]. As resolution is halved and area is squared, the amount of processing is exponentially small. The purpose of PBB is to

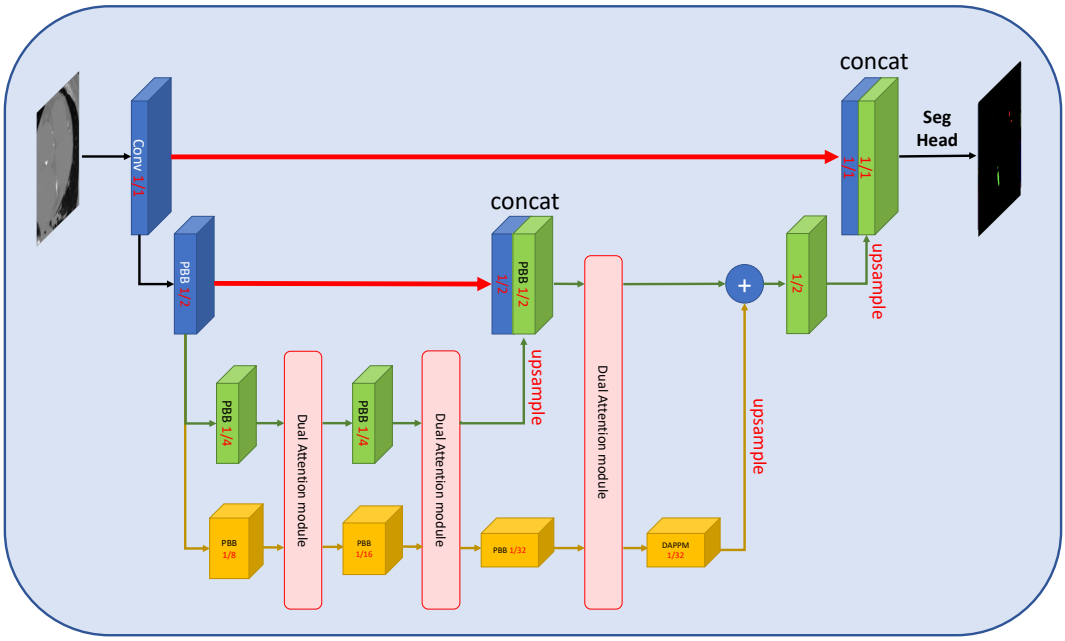

Fig. 1. DDR-SAPC-U schematic diagram Changes from DDRNet are shown in red. The stride of the convolution is reduced and the feature resolution is increased.

effectively extract features using multiple resolutions while the increase in processing amount is maintained to a small amount.

IV. Experiment

A. Dataset

In this paper, we used a cardiac CT coronary artery calcification data set to verify that the proposed method can be applied to medical images depicted in Figure 3. This dataset was created for detecting coronary artery calcification [23] [24] [25], which is considered to be highly related to coronary artery disease. We manually performed the annotation. The objects of label are the left anterior descending artery and calcifications present in three coronary arteries: left anterior descending artery (LAD, LMT), left circumflex artery (LCX), and right coronary artery (RCA) In this dataset, there are four classes including the class without calcification. The number of training verification data is 5600 (obtained from 100 patients), and the number of test data is 1240 (obtained from 21 patients).

B. Pretreatment and training method

The coronary artery calcification dataset was not preprocessed during training. The CT image was input as is. SGD was used as the optimization method, and momentum and weight decay were set to 0.9 and 0.0005. The initial value of the learning coefficient was 0.01, and training was performed with a batch size of 4 and an epoch number of 50. OHEM CE loss was the loss function.

The model was evaluated by inputting test data for the coronary artery calcification data set and examining the detection accuracy and detection speed, computational load, and the number of parameters. The mean Intersection over Union (mean IoU) was considered as an evaluation index for detection accuracy.

C. Visualization of detection results at coronary artery calcification dataset

Visualization example of the detection results in the coronary artery calcification test data set shows as below. Figure 3(b) shows the classification of classes in the correct and output images. From the upper left of the figure, the input image, the output of UNet, the output of UNet++, and from the lower left the correct image, the output of TransUNet, and the output of DDR-SAPC-U. First, in Figure 4(a), UNet, UNet++, and TransUNet incorrectly detect aortic calcification as RCA calcification for a correct image with LAD and LCX calcification. In comparison, DDR-SAPC-U improved this false detection. Furthermore, LCX calcification was also detected in a form close to the correct image. Figure 4(b) displays an output diagram for images with LAD, LCX, and RCA calcifications. UNet and UNet++ failed to detect LCX calcification, however TransUNet and DDR-SAPC-U detected it correctly.

D. Experiment results

1) Testing Environment: For the proposed models, all the training and testing pipelines, as well as baselines were implemented using PyTorch 1.6.0+cu92 framework in a Python 3.8.5 virtual environment. The graphics processing unit used in the training pipeline was NVIDIA Quadro P6000, and the RAM was 24 GB. Jupyter notebooks were utilized to conduct the experiments.

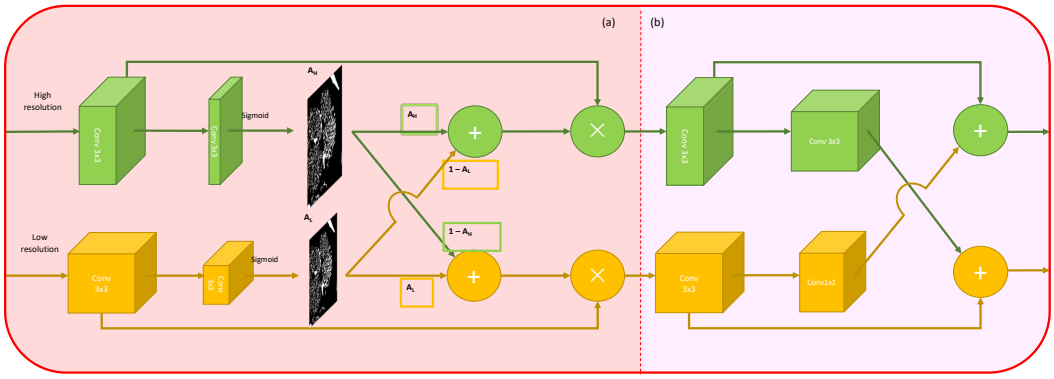

Fig. 2. Dual Attention Module. Learn more about the Dual Attention Module. This method adds the structure of part (a) to part (b) of the conventional method. Part (a) generates spatial attention at both high and low resolutions for the purpose of emphasizing various objects.

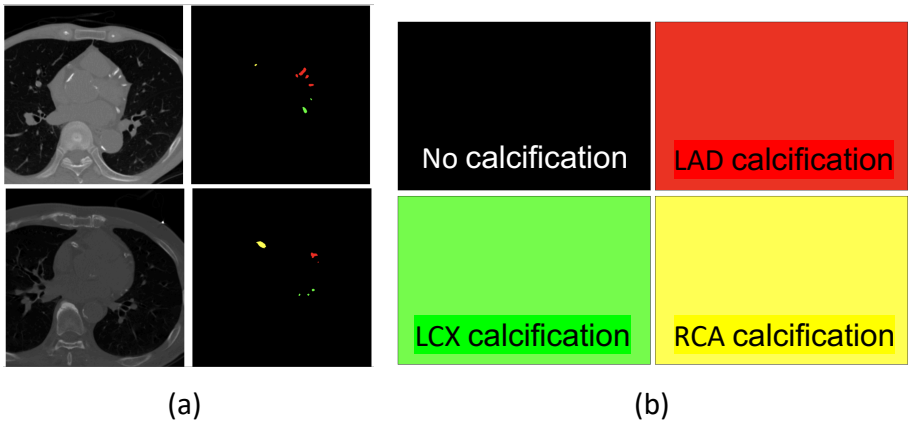

Fig. 3. Coronary Artery Calcification Dataset Image Example. (a) Left side: Input images. Right side: Correct images.(b) Coronary calcification dataset color map. Black indicates no calcification, red indicates LAD calcification, green indicates LCX calcification, and yellow indicates RCA calcification class.

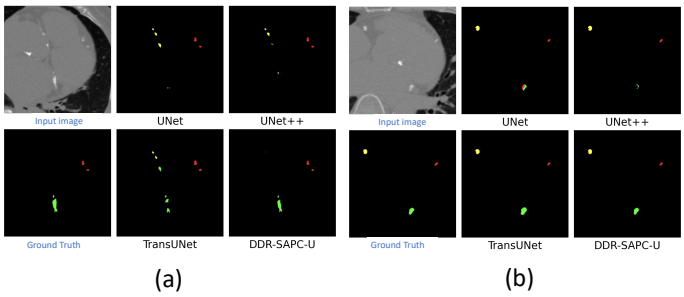

Fig. 4. Output images by model in coronary artery calcification dataset Black indicates no calcification, red indicates LAD calcification, green indicates LCX calcification, and yellow indicates RCA calcification class.

2) Experiment results: Using the proposed method DDR-SAPC-U, we conducted experiments according to the method described above. We added UNet [14], UNet++ [15], and TransUNet [16] as conventional methods for semantic segmentation tasks in the medical field and comparatively evaluated them.

The results are listed in Table I. The proposed method recorded high accuracy in all classes compared to other methods. FLOPs were relatively low values, indicating that the proposed method is superior in terms of accuracy, computational complexity, and lightness.

V. Discussion

DDR-SAPC-U recorded relatively fewer FLOPs and maximum accuracy in the Table I . In particular, the recognition accuracy has improved in the calcification classes located in the LAD and LCX. These calcifications tend to be difficult to recognize, especially at the branching points of the coronary arteries. The feature quantity was effective in the spatial direction by DAM. RCA improved accuracy less than that other methods. This is probably because of the shallow convolutional layers and the lightweight structure of the proposed method, DDR-SAPC. Therefore, the accuracy can be improved by adding more channels in the convolution and deepening the convolution layers.

VI. Conclusion

This research proposed a fast segmentation method, DDR-SAPC, which adds spatial attention and convolution with low resolution. Compared with UNet, UNet++, and TransUNet,

TABLE I
CORONARY ARTERY CALCIFICATION TEST DATASET RESULTS

Architecture	IoU by class				mIoU	GFLOPs	Params
	no calcification	LAD	LCX	RCA			
Unet [14]	99.99	80.74	52.55	73.28	76.64	**54.62**	**7.76M**
Unet++ [15]	99.99	76.58	55.97	80.61	78.29	137.95	9.16M
TransUnet [16]	99.98	66.41	42.83	63.67	68.23	62.19	20.07M
DDR-SAPC-U (Our proposed model)	99.99	**88.33**	**61.64**	**80.65**	**82.65**	60.15	13.42M

which are conventional methods in the medical field, the proposed method demonstrated the highest accuracy. Furthermore, it is relatively lightweight in terms of computational complexity and has been validated to be effective in the medical field as well. And we contribute that the proposed model is effective for segmentation in a wide range of segmentation applications. Future tasks include adding channel direction attention to the network, improving the accuracy of difficult classes such as walls and fences, and changing the network structure.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant for Toyohashi Heart Center Smart Hospital Joint Research Course and the Grant-in-Aid for Scientific Research (C) (issue numbers 22K12149 and 22K12040) and Knowledge Hub Aichi in Priority Research Project DX "Realization of a smart hospital that brings together IT and AI technologies".

REFERENCES

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *CoRR*, vol. abs/1908.07919, 2019. [Online]. Available: http://arxiv.org/abs/1908.07919

[3] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *CoRR*, vol. abs/2101.06085, 2021. [Online]. Available: https://arxiv.org/abs/2101.06085

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[6] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. [Online]. Available: http://arxiv.org/abs/1612.01105

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: http://arxiv.org/abs/1511.00561

[9] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *CoRR*, vol. abs/1808.00897, 2018. [Online]. Available: http://arxiv.org/abs/1808.00897

[10] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation," *CoRR*, vol. abs/2004.02147, 2020. [Online]. Available: https://arxiv.org/abs/2004.02147

[11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016. [Online]. Available: http://arxiv.org/abs/1606.02147

[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," *CoRR*, vol. abs/1704.08545, 2017. [Online]. Available: http://arxiv.org/abs/1704.08545

[13] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong, "Semantic flow for fast and accurate scene parsing," *CoRR*, vol. abs/2002.10120, 2020. [Online]. Available: https://arxiv.org/abs/2002.10120

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018. [Online]. Available: http://arxiv.org/abs/1807.10165

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[18] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *CoRR*, vol. abs/1809.02983, 2018. [Online]. Available: http://arxiv.org/abs/1809.02983

[19] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," *CoRR*, vol. abs/2107.00782, 2021. [Online]. Available: https://arxiv.org/abs/2107.00782

[20] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *CoRR*, vol. abs/1909.11065, 2019. [Online]. Available: http://arxiv.org/abs/1909.11065

[21] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *CoRR*, vol. abs/2005.10821, 2020. [Online]. Available: https://arxiv.org/abs/2005.10821

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[23] M. J. Budoff and K. M. Gul, "Expert review on coronary calcium," *Vascular Health and Risk Management*, vol. 4, no. 2, pp. 315–324, 2008, pMID: 18561507. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/vhrm.s1160

[24] S. Matsuoka, T. Yamashiro, A. Diaz, R. S. J. Estépar, J. C. Ross, E. K. Silverman, Y. Kobayashi, M. T. Dransfield, B. J. Bartholmai, H. Hatabu, and G. R. Washko, "The relationship between small pulmonary vascular alteration and aortic atherosclerosis in chronic obstructive pulmonary disease: quantitative CT analysis," *Acad. Radiol.*, vol. 18, no. 1, pp. 40–46, Jan. 2011.

[25] M. J. Blaha, J. Yeboah, M. Al Rifai, K. Liu, R. Kronmal, and P. Greenland, "Providing evidence for subclinical cvd in risk assessment," *Global Heart*, vol. 11, no. 3, pp. 275–285, 2016, legacy of Multiethnic Study of Atherosclerosis. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211816016307086