ViTAL-CT: Vision Transformers for High-Risk Plaque Classification in Coronary CTA

Anjie Le[1⋆], Jin Zheng[1,2⋆], Tan Gong[3], Quanlin Sun[4], Jonathan Weir-McCall[5], Declan P. O'Regan[2], Michelle C. Williams[6], David E. Newby[6], James H.F. Rudd[1], and Yuan Huang[1,4,7⋆⋆]

[1] Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, UK
[2] MRC Laboratory of Medical Sciences, Imperial College London, UK
[3] School of Biomedical Engineering, Tsinghua University, Beijing, China
[4] Department of Radiology, University of Cambridge, UK
[5] Department of Cardiovascular Imaging, School of Biomedical Engineering and Imaging Sciences, King's College London, UK
[6] British Heart Foundation Centre of Research Excellence, University of Edinburgh, UK
[7] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

yh288@cam.ac.uk

Abstract. High-risk plaque (HRP) detected by coronary CT angiography (CTA) is associated with increased risks of major adverse cardiovascular events such as heart attack. Current identification of HRP characteristics involves labor-intensive segmentation of plaques, requiring substantial time and expert knowledge. In this work, we propose a novel coronary cross-sectional Vision Transformer (ViT) framework that bypasses the need for explicit segmentation by directly predicting the presence of HRP. Our approach extracts cross-sectional slices along the coronary centerline, ensuring that the model focuses on the artery. By leveraging the standard patch-based input of ViT, we capture not only the coronary cross-section itself but also surrounding contextual information (e.g., adipose tissue). Furthermore, we incorporate multiple levels of detail by combining the cross-sections from proximal and distal positions with their corresponding CTA axial planes, forming a comprehensive cross-sectional representation. We also embedded the actual 3D position of each cross-section into the positional encoding of the Transformer to enhance spatial awareness. Experimental results of 3,068 coronary arteries demonstrate that our method outperforms conventional approaches, highlighting its potential to optimize clinical decision-making in the care of coronary artery diseases[1].

Keywords: Vision Transformer (ViT) · Coronary CT Angiography (CTA) · High-Risk Plaque (HRP).

⋆ Equal contribution.
⋆⋆ Corresponding author.
[1] https://github.com/JZCambridge/ViTAL-CT-MICCAI25

1 Introduction

Coronary artery disease (CAD), primarily caused by atherosclerosis, is the leading cause of mortality and morbidity worldwide. Coronary computed tomography angiography (CTA) has emerged as a first-line imaging modality to assess coronary atherosclerotic plaques and guide CAD treatment. Independent of conventional cardiovascular risk factors, CTA-depicted high-risk plaque (HRP) features are associated with an increased risk of major adverse cardiovascular events (MACEs), such as death and myocardial infarction. However, clinical translation of CTA HRP-based risk stratification is limited due to technical challenges: (a) Manual identification of HRP requires substantial expertise and is prone to human error [1]; (b) Atherosclerotic plaques can be focal or diffuse, complicating lesion alignment and HRP localization; (c) The small size of coronary arteries compared to the entire CTA volume makes conventional deep learning pipelines computationally inefficient.

Both radiomics and deep learning approaches have been applied to identify HRPs. The radiomics approach typically utilizes handcrafted histogram and texture features, as well as pre-defined plaque characteristics such as stenosis and plaque burden. These features are then fed into a statistical learning framework for variable selection and classification. Radiomics has been shown to outperform conventional clinical features in the detection of the napkin ring sign [2]. Using the SCOT-HEART dataset, eigen radiomics were found to add predictive value for future infarctions [3].

The radiomics approach is computationally less demanding, but its predefined features are highly dependent on vendors and scan settings. Furthermore, it requires a separate segmentation pipeline for the lumen and plaque wall, which further affects the performance and robustness of the analysis [4]. Recently, deep learning approaches have been explored to mitigate these drawbacks. For instance, a study used a hierarchical convolutional long short-term memory (ConvLSTM) network to segment plaque components and calculate plaque characteristics [5]. Another pipeline, Coronary R-CNN, was developed for automated plaque analysis. Inspired by Faster R-CNN, it includes an object detection module to localize the diseased segment and a multi-head analysis module to calculate stenosis [6]. However, there is limited work on deep learning approaches for HRP features, possibly due to previous reliance on object detection.

In this work, we propose a novel end-to-end approach for classifying HRPs by directly analyzing cross-sectional slices along the coronary centerline. Our method classifies HRP collectively, consistent with clinical studies [7] linking aggregated HRP burden to increased MACE risk. This eliminates the need for labor-intensive plaque or vessel wall segmentation, improving scalability. Our method leverages the patch-based architecture of Vision Transformers (ViTs), treating each cross-section as a patch to preserve anatomical continuity while capturing localized plaque features and contextual relationships across slices. We also fuse multi-scale representations by combining proximal and distal cross-sectional planes with a coronary-tailored ConvNeXt block. This enhances the

model's ability to capture diffuse plaque patterns, while the U-Net bottleneck integrates comprehensive cross-sectional information. Furthermore, we embed the 3D positional coordinates of each cross-section into the transformer's positional encoding, improving spatial awareness. Our proposed framework, **ViTAL-CT** (Vision Transformers for High-Risk Plaque Classification in Coronary CTA), eliminates the need for manual annotations and outperforms conventional 2D/3D CNN methods. Our key contributions are:

- The first segmentation-free ViT framework for HRP classification, analyzing centerline-aligned cross-sections directly, tailored for CTA.
- A hybrid multi-scale representation combining proximal/distal cross-sections, axial context, and 3D geometry to improve sensitivity to diffuse plaques.
- Large-scale validation on 3,068 coronary arteries, demonstrating state-of-the-art performance for HRP detection.

2 Methodology

Our proposed ViTAL-CT framework integrates a ViT, a ConvNeXt block, and a U-Net bottleneck to classify HRPs in coronary CTA (Fig. 1). This hybrid architecture processes three complementary input streams to capture both local and global features of coronary plaques:

- **Cross-Section Stream**: A grayscale 2D cross-section centered on the coronary artery, providing localized information of the region of interest.
- **Multi-Slice Context Stream**: A coronary-tailored ConvNeXt block applied to 9 adjacent slices (spanning proximal and distal slices) to capture longitudinal plaque morphology and improve contextual understanding.
- **Global Context Stream**: A U-Net bottleneck layer that encodes the overall axial plane context around the coronary artery, providing comprehensive spatial information.

The use of global and multi-slice contexts is intended to mirror clinical reasoning, in which plaques—such as positive remodeling, soft plaque, and mixed composition—require examining both the proximal and distal segments of the artery. All above streams are fused into a unified 3-channel input, where each ViT patch token represents a coronary cross-section enriched with both local (from the cross-section) and global (from multi-slice and axial context) information, extracted features from the same coronary location. Additionally, we embed 3D positional coordinates into the Transformer's positional encoding to ensure spatial awareness across slices. Coronary patch-specific augmentation and masking techniques are also adapted to further improve model performance [8].

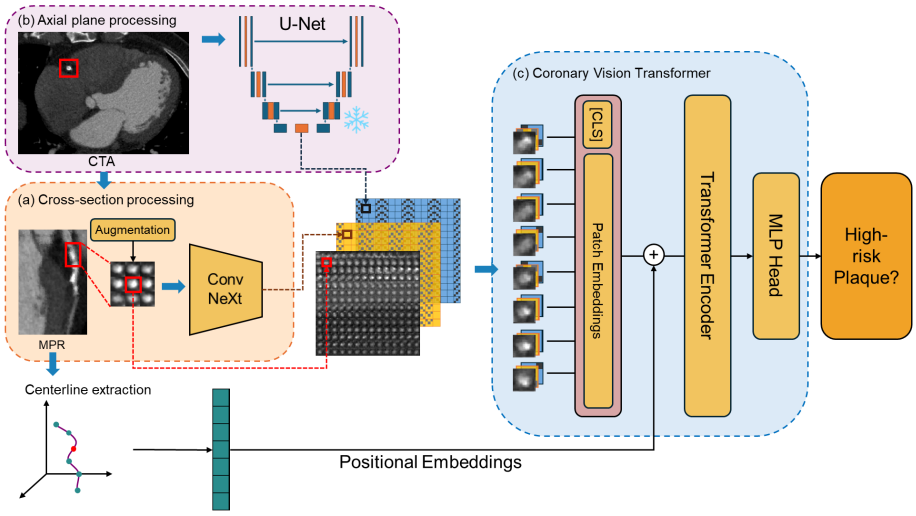

Fig. 1: ViTAL-CT framework. Three parallel streams process the central cross-sections, adjacent slices (ConvNeXt) (a), and axial contexts (UNet) (b). These streams are fused into 3-channel tokens with 3D positional embeddings. A ViT backbone (c) processes the tokens to predict high-risk plaque, guided by position-aware self-attention.

2.1 Coronary-tailored ConvNeXt block for Multi-Slice Context

To better integrate the longitudinal plaque morphology in HRP identification and improve the model's understanding of context, each 2D coronary cross-sectional slice is concatenated with its 8 adjacent slices to form a new patch, denoted $P_{\text{cross}} \in \mathbb{R}^{H \times W}$, which contains multilayer semantic information. Inspired by ConvNeXt [9], we proposed a coronary-tailored ConvNeXt block (CTCB) to reduce the dimensionality of P_{cross} and extract features with minimal computational resource consumption. The CTCB first performs dimensionality reduction using a 3×3 convolution with a stride of 3, and padding of 1. Given the tubular structure of the coronary arteries oriented along the tangential axis, an anisotropic depth-wise convolution with a kernel size of 7×3 is followed in CTCB to emphasize the anisotropic nature of the plaque. A vessel-aware attention is also applied to amplify the difference between calcified and low-attenuated regions of the plaque. Finally, a residual connection is adopted to keep abundant context information from the early layers [10]. After dimension reduction, the D layers of the CTCB outputs, denoted as $P_{\text{MSC}} \in \mathbb{R}^{\frac{H}{3} \times \frac{W}{3}}$, are then ordered from proximal to distal to obtain multilayer context channels for ViT input with a size of $(\sqrt{D} \times \frac{H}{3}, \sqrt{D} \times \frac{W}{3})$. For coronary CTA, $H = W = 48$ and $D = 196$ were used.

2.2 U-Net for Global Context

To capture comprehensive spatial information from the axial plane, we trained a U-Net [11] on 400 manually segmented arteries as a feature extractor, among which 320 arteries were used for training and 80 for validation. The arteries were selected at the patient level to avoid leakage. For each artery, 10 slices were sampled at equal distances to be included in the process. Testing was performed separately on all slices on 4 additional arteries.

The U-Net structure consists of an encoder path with five downsampling blocks and the corresponding decoder blocks with skip connections. Each encoder block comprises a max-pooling operation for downsampling, followed by two convolutional layers with ReLU activation. We also employed batch normalization and a dropout rate of 0.1 to prevent overfitting.

The bottleneck is designed to have a size of $16 \times 16 \times 256$, whose feature is then extracted by adaptive pooling from each channel and used as one of the input streams for the ViT model, as a compact representation of the global context for HRP identification. In our experiment, the encoder of a U-Net of testing dice score 0.72 was used for the static feature extraction process.

2.3 Vision Transformer & Positional Embedding

For classification, we use a ViT base model with a patch size of 16, an embedding dimension of 768, a depth of 12, and 12 attention heads pre-trained on ImageNet-21k [12]. Low-rank adaptation (LoRA) is applied to avoid overfitting from fine-tuning all parameters [13]. The Transformer architecture allows for long-range dependency which is more suitable for atherosclerotic plaque assessment, as the lesions are usually diffuse. The 16×16 patch, with each pixel corresponding to 0.35 mm, sufficiently covers the average coronary diameter of approximately 3.5-4 mm. It also captures surrounding contextual information such as adipose tissue, an important indicator of coronary inflammation [14]. In the positional encoding, we further incorporate the absolute spatial location information of the coronary artery segment, centering 4 slices proximal and distal around the cross-section of interest, to provide not only the spatial location but also the local geometry.

2.4 Augmentation

To enhance model robustness and address the limited availability of coronary CTA data, we employ coronary-specific augmentations that preserve anatomical integrity while introducing controlled variations. Augmentations are applied across entire arterial structures to maintain spatial consistency across slices [15].

- **Spatial Transformations**:
  - Random translations (± 2 pixels) along the height and width axes, orthogonal to the centerline, to simulate minor positional shifts.
  - Selective flipping along anatomical axes (height, width, and diagonal) to introduce plausible variations in vessel orientation.

These transformations promote invariance to minor anatomical differences while preserving the core vascular structure.

- **Rotational Augmentation**: Since coronary arteries can appear in diverse orientations, we apply random in-plane rotations of up to 90 degrees along the vessel's longitudinal axis. This enhances the model's ability to generalize across different imaging perspectives.
- **Masked Region Learning**: To encourage robust feature learning, we randomly mask 20% of coronary cross-sections, following self-supervised learning paradigms such as Masked Autoencoders (MAE) and Masked Language Modeling (MLM) [8,16]. This forces the model to infer missing structural details, improving contextual understanding and resilience to incomplete imaging data.

All augmentations are calibrated to ensure clinical validity while introducing sufficient variability to enhance model generalization.

3 Experiments and Results

3.1 Dataset and Experimental Setup

A total of 1,060 patients from the SCOT-HEART study [17] were included. From each case, three main coronary arteries (RCA, LAD, and LCX), consisting of proximal, middle, and distal segments (for a total of 3,068 segments), were extracted. An artery is defined to contain high-risk plaque when any plaque contains one or more of the following characteristics in the proximal or middle segment: positive remodeling, low-attenuation plaque, spotty calcification, or napkin-ring sign [7]. Overall, 675 diseased arteries were identified as having high-risk plaque.Coronary cross-sections were generated along the vessel centerline at 0.5mm intervals, preserving the same in-plane resolution as the original CTA data. Figure 2 shows an example of an LAD containing high-risk plaque, along with illustrations of cross-sections and augmented inputs.

The dataset was split into training (60%), validation (20%), and test (20%) sets. All dataset splits were performed at the patient level to prevent data leakage across training, validation, and test sets. Consequently, other models for comparison and ablation are evaluated on the exact same patient set. Stratified sampling and coronary-specific data augmentation were employed to maintain a balanced class distribution during training. All experiments were conducted using PyTorch on an NVIDIA RTX 4090 (24 GB). We used the AdamW optimizer with a batch size of 32 and an initial learning rate of 0.0001. The F1 score was our primary metric of interest so the best model was chosen based on the highest F1 score in validation. We also reported precision, sensitivity (recall), and area under the receiver operating characteristic curve (AUC).

3.2 Comparison with Other Methods

We compared our proposed model with various ViT configurations of different sizes, as well as with convolution-based ResNet50 models, clinically popular mod-

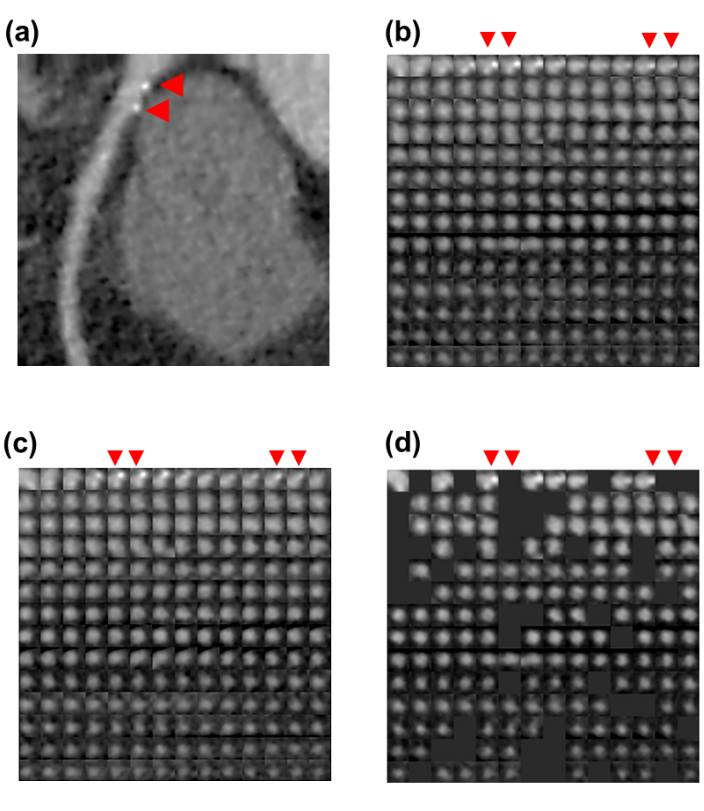

Fig. 2: Example of a high-risk plaque (spotty calcification). (a) Curved multiplanar reconstruction with the plaque indicated by an arrowhead. (b) Coronary cross-sections. (c) Cross-sections with augmentations. (d) Cross-sections with masking.

els in coronary CT studies [18, 19], using 2D cross-sections and 3D straightened multiplanar reconstructions (Table 1) [10, 12].

In general, larger ViT models tended to perform worse on this relatively small dataset, reflecting the known data-hungry nature of Transformers. Notably, incorporating LoRA for parameter-efficient fine-tuning on medical images yielded performance gains. The ViT Base & LoRA achieved a 6% increase in AUC compared to ViT Tiny. Meanwhile, ResNet50 2D outperformed all standard ViT variants, likely due to the importance of local feature extraction in coronary cross-sections, aligning with the popularity of CNN-based approaches in medical imaging.

By integrating a ConvNeXt block for local feature enhancement, our **ViTAL-CT** model surpassed all other methods, yielding a 7% improvement in AUC and a 4% improvement in F1 score over the strongest baseline. This suggests that

combining convolution-based local feature extraction with ViT's global context captures both fine-grained plaque textures and broader anatomical variations in coronary arteries, yielding a better trade-off between precision and recall.

Method	AUC	Precision	Recall	F1
ViT Tiny	0.755	0.678	0.710	0.690
ViT Small	0.743	0.664	0.701	0.675
ViT Base	0.742	0.658	0.689	0.668
ViT Base & LoRA	0.799	0.685	0.735	0.698
ResNet50 2D	0.806	0.721	0.697	0.707
ResNet50 3D	0.774	0.699	0.707	0.703
ViTAL-CT (Ours)	**0.818**	**0.735**	**0.771**	**0.749**

Table 1: Performance comparison of four ViT-based models (varying sizes and LoRA usage) and ResNet50 models (2D vs. 3D cross-sections). Our ViTAL-CT model achieves the highest scores across all metrics.

3.3 Ablation Experiments

Ablation results (Table 2) validate that each module, including U-Net global context, positional embedding, ConvNeXt multi-slice stream, and data augmentation, contributes uniquely to performance, supporting our hybrid architecture design.

Method	AUC	Precision	Recall	F1
Without U-Net	0.758	0.692	0.705	0.698
Using Residual Block	0.685	0.620	0.620	0.620
Without Positional Embedding	0.775	0.694	0.732	0.707
Without Augmentation	0.737	0.649	0.680	0.658
ViTAL-CT (Ours)	**0.818**	**0.735**	**0.771**	**0.749**

Table 2: Ablation study of ViTAL-CT components. Removing the ConvNeXt block, global U-Net features, or positional embeddings each leads to a measurable reduction in overall performance.

Replacing the ConvNeXt block with a standard residual block causes roughly a 13% decrease in both AUC and F1, highlighting the importance of the ConvNeXt architecture in capturing coronary-specific details and fine-grained plaque textures. The data-hungry nature of the Transformer architecture and the large variations of plaque and vessel imaging in real-world clinical settings can be observed by the drop in performance by 10% in AUC and 12% in F1 when there is

no augmentation. Removing the global features from the U-Net bottleneck also has a clear impact, with decreases of 6% in AUC and 5% in F1. This likely comes from losing broader context regarding other arterial segments, which is crucial given the diffuse and systemic nature of atherosclerosis. Finally, removing the 3D spatial positional embedding yields a smaller yet noticeable drop of about 4% in both AUC and F1, suggesting that spatial awareness still provides a beneficial signal, even though the limited size of this coronary dataset constrains the network's ability to learn more sophisticated 3D geometric relationships.

4 Conclusion

In this paper, we propose a novel segmentation-free and multi-scale ViT framework tailored for coronary CTA plaque assessment. Evaluation on 3,068 coronary arteries from 1,060 patients demonstrated that the presented method achieved better results than the state-of-the-art methods. Future work includes better incorporating the arterial positional embeddings, adding non-imaging clinical features, and validating the pipeline in additional datasets.

5 Disclosure of Interests

The SCOT-HEART trial was funded by The Chief Scientist Office of the Scottish Government Health and Social Care Directorates (CZH/4/588). MCW is supported by British Heart Foundation (FS/ICRF/20/26002). JHFR and YH are part-supported by BHF CRE(RE/24/130011) and EPSRC. JHFR is part-supported by NIHR Cambridge BRC and Wellcome Trust.

References

1. Ihdayhid, A.R., Sehly, A., Lan, N.S.R., Denston, N., Chow, B.J.W., Newby, D.E., Williams, M.C., Dwivedi, G.: Characterising high-risk plaque on cardiac ct. Journal of Medical Imaging and Radiation Oncology (2024)
2. Kolossváry, M., Karády, J., Szilveszter, B., Kitslaar, P., Hoffmann, U., Merkely, B., Maurovich-Horvat, P.: Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign. Circulation: Cardiovascular Imaging **10**(12), e006843 (2017)
3. Kolossváry, M., Lin, A., Kwiecinski, J., Cadet, S., Slomka, P.J., Newby, D.E., Dweck, M.R., Williams, M.C., Dey, D.: Coronary plaque radiomic phenotypes predict fatal or nonfatal myocardial infarction: Analysis of the scot-heart trial. JACC: Cardiovascular Imaging (2024)
4. Kolossváry, M., De Cecco, C.N., Feuchtner, G., Maurovich-Horvat, P.: Advanced atherosclerosis imaging by ct: Radiomics, machine learning and deep learning. Journal of Cardiovascular Computed Tomography **13**(5), 274–280 (2019)
5. Lin, A., Manral, N., McElhinney, P., Killekar, A., Matsumoto, H., Kwiecinski, J., Pieszko, K., Razipour, A., Grodecki, K., Park, C., Otaki, Y., Doris, M., Kwan, A.C., Han, D., Kuronuma, K., Flores Tomasino, G., Tzolos, E., Shanbhag, A.,

Goeller, M., Marwan, M., Gransar, H., Tamarappoo, B.K., Cadet, S., Achenbach, S., Nicholls, S.J., Wong, D.T., Berman, D.S., Dweck, M., Newby, D.E., Williams, M.C., Slomka, P.J., Dey, D.: Deep learning-enabled coronary ct angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. The Lancet Digital Health **4**(4), e256–e265 (2022)

6. Zhang, Y., Ma, J., Li, J.: Coronary r-cnn: Vessel-wise method for coronary artery lesion detection and analysis in coronary ct angiography. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 207–216. Springer Nature Switzerland, Cham (2022)
7. Williams, M.C., Moss, A.J., Dweck, M., Adamson, P.D., Alam, S., Hunter, A., Shah, A.S., Pawade, T., Weir-McCall, J.R., Roditi, G., van Beek, E.J., Newby, D.E., Nicol, E.D.: Coronary Artery Plaque Characteristics Associated With Adverse Outcomes in the SCOT-HEART Study. Journal of the American College of Cardiology **73**(3), 291–301 (2019)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021), https://arxiv.org/abs/2111.06377
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
10.
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
14. Antonopoulos, A.S., Sanna, F., Sabharwal, N., Thomas, S., Oikonomou, E.K., Herdman, L., Margaritis, M., Shirodaria, C., Kampoli, A.M., Akoumianakis, I., Petrou, M., Sayeed, R., Krasopoulos, G., Psarros, C., Ciccone, P., Brophy, C.M., Digby, J., Kelion, A., Uberoi, R., Anthony, S., Alexopoulos, N., Tousoulis, D., Achenbach, S., Neubauer, S., Channon, K.M., Antoniades, C.: Detecting human coronary inflammation by imaging perivascular fat. Science Translational Medicine **9**(398) (2017)
15. Liu, Y., Tian, Y., Wang, C., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Translation consistent semi-supervised segmentation for 3d medical images. IEEE Transactions on Medical Imaging (2024)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 4171–4186 (June 2019)
17. The SCOT-HEART Investigators: Coronary ct angiography and 5-year risk of myocardial infarction. New England Journal of Medicine **379**(10), 924–933 (2018)
18. Mu, D., Bai, J., Chen, W., Yu, H., Liang, J., Yin, K., Li, H., Qing, Z., He, K., Yang, H.Y., Zhang, J., Yin, Y., McLellan, H.W., Schoepf, U.J., Zhang, B.: Calcium

scoring at coronary ct angiography using deep learning. Radiology **302**(2), 309–316 (2022). https://doi.org/10.1148/radiol.2021211483, https://doi.org/10.1148/radiol.2021211483, pMID: 34812674

19. Tan, E.W.P., Cheng, N., Leng, S., Baskaran, L., Teo, L., Yew, M.S., Singh, M., Go, M.C.R., Huang, W.M., Raffiee, N.A.S., Chan, M.Y.Y., Ngiam, K.Y., Tan, S.Y., Lee, H.K., Zhong, L, A.I.: Performance of modified resnet model for calcium score from non-contrast computed tomography. European Heart Journal **44**(Supplement2) (11 2023). https://doi.org/10.1093/eurheartj/ehad655.2927, https://doi.org/10.1093/eurheartj/ehad655.2927