A Hybrid CNN-Transformer Feature Pyramid Network for Granular Abdominal Aortic Calcification Detection from DXA Images

Zaid Ilyas[1,2], Afsah Saleem[1,2], David Suter[1,2], John T. Schousboe[4], William D. Leslie[5], Joshua R. Lewis[1,2], and Syed Zulqarnain Gilani[1,2,3]

[1] Centre for AI & ML, School of Science, Edith Cowan University, Australia
[2] Nutrition and Health Innovation Research Institute, Edith Cowan University, Australia
[3] Computer Science and Software Engineering, The University of Western Australia
[4] Park Nicollet Clinic and HealthPartners Institute, Minneapolis,USA
[5] Department of Medicine and Radiology, University of Manitoba, Canada
z.ilyas@ecu.edu.au

Abstract. Cardiovascular Diseases (CVDs) stand as the primary global cause of mortality, with Abdominal Aortic Calcification (AAC) being a stable marker of these conditions. AAC can be observed in Dual Energy X-ray absorptiometry (DXA) lateral view Vertebral Fracture Assessment (VFA) scans, usually performed for the detection of vertebral fractures. Early detection of AAC can help reduce the risk of developing clinical CVD by encouraging preventive measures. Recent efforts to automate DXA VFA image analysis for AAC detection are restricted to either predicting an overall AAC score, or they lack performance in granular AAC score prediction. The latter is important in helping clinicians predict CVD associated with the diminished Windkessel effect in the aorta. In this regard, we propose a hybrid Feature Pyramid Network (FPN) based CNN-Transformer architecture (Hybrid-FPN-AACNet) that employs a novel Dual Resolution Self-Attention (DRSA) mechanism to enhance context for self-attention by working on two different resolutions of the input feature map. Moreover, the proposed architecture also employs a novel Efficient Feature Fusion Module (EFFM) that efficiently combines the features from different hierarchies of Hybrid-FPN-AACNet for regression tasks. The proposed architecture has achieved State-Of-The-Art (SOTA) performance at a granular level compared to previous work. The code is available at https://github.com/zaidilyas89/Hybrid-FPN-AACNet.

Keywords: Hybrid-FPN-AACLiteNet · Dual Resolution Self-Attention · Efficient Feature Fusion Module.

1 Introduction

Cardiovascular Diseases (CVDs) are the leading cause of death worldwide, affecting 17.9 million people each year [1]. Atherosclerosis, a precursor to CVDs, causes calcification in blood vessels, with the abdominal aorta being one of the initial sites where this condition manifests [2,3]. Abdominal Aortic Calcification (AAC) is a stable marker of atherosclerosis and can help predict future CVD events [4,5,6,7]. Early detection of AAC can help promote preventive measures to mitigate adverse outcomes related to CVD, including premature death. AAC can be detected using different imaging modalities such as Computed Tomography (CT) [8] and digital X-ray imaging [9], however, Dual-Energy X-ray Absorptiometry (DXA) is the modality of choice given its low-cost and low-radiation exposure [5,6,9,10]. However, identification of AAC from DXA VFA poses challenges due to low resolution, potential artifacts, and poorly delineated vertebral boundaries.

AAC-24 scoring [11] is a semi-quantitative method to measure calcification in the aorta parallel to vertebrae L1-L4. Expert radiologists read the scans and score the calcification on the anterior and posterior walls of each lumbar vertebrae L1-L4. Although the process is granular in nature, clinicians mainly use the overall AAC-24 score, which is the sum of the calcification detected in the aorta. Manual scoring of AAC-24 is laborious, time-consuming, and costly. Prior research efforts [12,13,14,16,17,18] to automate DXA VFA image analysis have focused mainly on detecting the overall AAC-24 score i.e. a single scalar output in the range of 0-24. Despite reporting satisfactory performance, these methods lack the capability to locate fine-grained calcification in the aortic regions adjacent to vertebrae L1 to L4. The significance of such granular scores cannot be underestimated. It is well known that elasticity in the aorta maintains the necessary Windkessel effect [19] throughout the cardiac cycle (systole and diastole). Calcification, on the other hand, stiffens the aorta, leading to elevated systolic blood pressure, left ventricular hypertrophy, and eventually congestive heart failure [20]. This loss of elasticity, particularly near the heart (i.e., near the L1 and L2 vertebrae region) increases pulse wave velocity [21,22], reducing the Windkessel effect. Therefore, identifying the location of AAC can help clinicians predict CVDs associated with diminished Windkessel effect in the aorta. The only work in granular AAC detection from VFA DXA images is of Gilani et al. [15], which utilizes an LSTM to sequentially predict these scores. However, it effectively lacks the ability to capture local fine-grained patterns and global long-range dependencies. Thus, [15] reports low agreement with ground truth. Consequently, we have designed the Hybrid-FPN-AACNet, a framework that introduces more explainability and accuracy in predicting granular-level AAC from DXA VFA images.

A drawback of existing deep learning approaches used for AAC detection from DXA images [14,15,16] is the reliance on the output from the last layer of the CNN backbone. Layers near the output end of the CNN backbone (deeper layers) give a coarse location of calcification and add global context by considering the curvature and shape of the spine, while the layers near the input end of the CNN-backbone (shallow/initial layers) provide fine local information with precise location, but lack global context. The intrinsic locality of convolutional operations hinders the ability of CNNs to model long-range dependencies while preserving spatial information in images effectively. Therefore, any analysis performed on the output of one hierarchical level may lead to missing information at other levels. A straightforward approach is to use Feature Pyramid Networks

(FPN) [23]. However, direct utilization of FPN lacks global attention among features and effective mechanism for feature fusion targeting tasks like regression.

Recently, hybrid CNN-Transformer architectures have shown remarkable performance in different medical image processing areas, exploiting benefits from both architectures [24,25,26]. Inspired by the hybrid approach, we propose a novel hybrid FPN-based CNN-transformer model, Hybrid-FPN-AACNet. It detects granular level AAC as regression scores and achieves SOTA performance. Conventionally, the quadratic computational cost of Self-Attention (SA) in a transformer block is tackled using window-based approaches; however, they limit the ability of SA to handle global relationships among features. To address this, we specifically design a Dual-Resolution Self-Attention (DRSA) block that incorporates a window-based SA mechanism at two different resolutions of the same input feature map to capture both, the narrower fine context and the broader coarse context. Additionally, we also propose a novel Efficient Feature Fusion Module (EFFM) which combines the feature maps of consecutive hierarchies efficiently by calculating transposed self-attention i.e. among channels of concatenated feature maps. Overall, the main contributions of this work are as follows:

- A Dual Resolution Self Attention (DRSA) block that captures both high and low-frequency components at different spatial contexts of the input feature map.
- An Efficient Feature Fusion Fusion Module (EFFM) to efficiently fuse features from different hierarchical levels of FPN and incorporate self-attention information among them.
- An FPN-based hybrid CNN transformer model (Hybrid-FPN-AACNet) that exploits the potentials of a feature pyramid network built by marrying a CNN backbone with a transformer architecture, and incorporating an efficient feature fusion strategy.

2 Proposed Framework

The architecture of our Hybrid-FPN-AACNet is shown in Fig.1. It comprises an FPN built on a CNN backbone, Transformer blocks employing DRSA, EFFM blocks to fuse features from consecutive hierarchies, and a regression head with eight outputs for the prediction of granular scores.

Dual Resolution Self-Attention (DRSA): DRSA is designed to achieve two objectives, i.e. to capture the information in diverse frequencies inherent in DXA VFA images and to expand the spatial context of windowed Multi-head Self-Attention (MSA). Different frequencies play distinct roles in encoding image patterns i.e. High-Frequency (HF) information deals with fine details like the texture of the object under consideration, and Low-Frequency (LF) information deals with the global structures, like the shape and curvature of the object. It is important to consider all these frequency components for a proper image analysis. Therefore, we apply DRSA on the feature maps of all hierarchical levels of FPN. DRSA splits the conventional MSA into two paths. One path employs

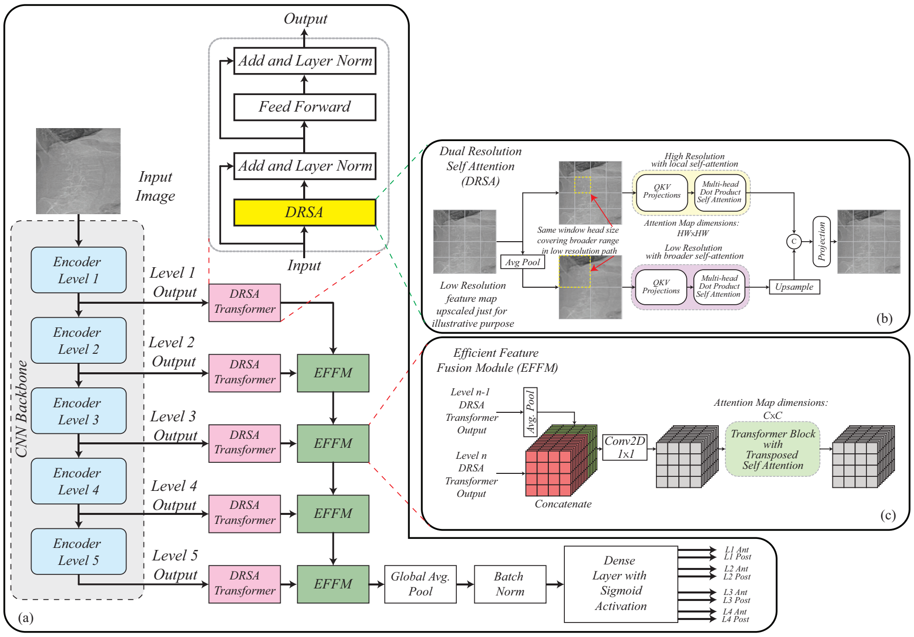
Output
Add and Layer Norm
Feed Forward
Add and Layer Norm
DRSA
Input
Input Image
CNN Backbone
Encoder Level 1
Encoder Level 2
Encoder Level 3
Encoder Level 4
Encoder Level 5
Level 1 Output
Level 2 Output
Level 3 Output
Level 4 Output
Level 5 Output
DRSA Transformer
EFFM
Global Avg. Pool
Batch Norm
Dense Layer with Sigmoid Activation
L1 Ant
L1 Post
L2 Ant
L2 Post
L3 Ant
L3 Post
L4 Ant
L4 Post
(a)
Dual Resolution Self Attention (DRSA)
High Resolution with local self-attention
QKV Projections
Multi-head Dot Product Self Attention
Attention Map dimensions: HW×HW
Same window head size covering broader range in low resolution path
Avg Pool
Low Resolution with broader self-attention
Upsample
C
Projection
Low Resolution feature map upscaled just for illustrative purpose
(b)
Efficient Feature Fusion Module (EFFM)
Level n-1 DRSA Transformer Output
Level n DRSA Transformer Output
Avg. Pool
Concatenate
Conv2D 1×1
Attention Map dimensions: C×C
Transformer Block with Transposed Self Attention
(c)

Fig. 1. (a) Hybrid-FPN-AACNet with Transformer blocks employing Dual Resolution Self Attention (DRSA) at each hierarchical level, and Efficient Feature Fusion Module (EFFM) combining features from consecutive hierarchies. (b) DRSA block performs self-attention at the actual and low-resolution versions of the feature map to capture fine-grained local context and coarse broader context. (c) EFFM - efficiently combines feature maps by calculating self-attention among channels of concatenated features.

local windowed-SA on an actual-sized input feature map to encode HF interactions, and the other path first downsamples the input feature map using average pooling (a low-pass filter operation) and then applies windowed-SA on it which captures LF information. We use the same-sized windows in both SA paths. The use of the same window size on a reduced-size feature map increases the spatial context of windowed SA and introduces overlap when the encoded feature maps are eventually combined. Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, where H, W, and C are the height, width, and the number of channels of the feature map respectively, the DRSA down-samples it using average pooling to generate a feature map $X' \in \mathbb{R}^{H/r \times W/r \times C}$, where r is the reduction factor e.g. 2. Separate Query, Key, and Value embeddings are generated for X and X' using linear transformation. Considering a single head, the self-attention for X, and X', abbreviated as SA$_h$ and SA$_l$ respectively are calculated as softmax $\left(\frac{Q_h K_h^T}{\alpha}\right) V_h$ and softmax $\left(\frac{Q_l K_l^T}{\alpha}\right) V_l$ respectively, where Q_h, K_h, V_h, Q_l, K_l, and V_l are the query, key, and value embeddings for feature maps X and X', with respective self-attentions SA$_h$ and SA$_l$. α is the learnable parameter to control SA.

$$\text{SA}_{total} = \text{Concat}(\text{SA}_h, \text{Upsample}(\text{SA}_l))W_p \tag{1}$$

SA$_{total}$ (the overall self-attention), is calculated by up-sampling the SA$_l$ using a learnable transposed convolution operation, concatenating it with SA$_h$, and

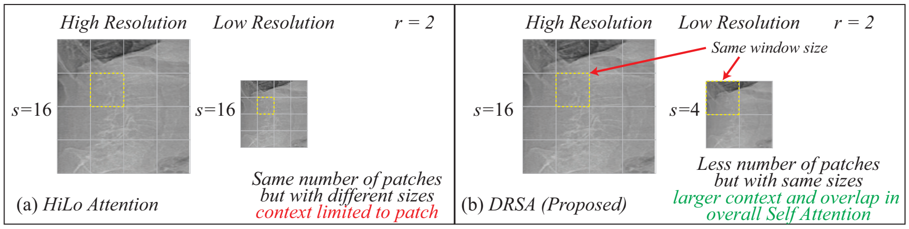

Fig. 2. (a) HiLo Attention [27] - Uses the same window count s in both the low and actual resolution SA paths which limits the context to individual patches. (b) DRSA (Our Proposed) - Uses the same window size in both resolution SA paths which increases the context in low-resolution SA path, and also introduces overlapping. In (a) and (b), r is the size reduction factor, and s is the window count.

then passing it through a linear layer W_p. Although DRSA is inspired by HiLo SA [27], we argue that the latter suffers from two drawbacks in the context of DXA image analysis. Firstly, since HiLO SA is designed for fast transformer models, it is restricted in spatial context. Secondly, it avoids overlap of patches while "paying attention". Fig. 2 shows the conceptual difference between DRSA and HiLo SA [27].

Efficient Feature Fusion Module (EFFM): For image regression, a naive integration approach is to concatenate the feature maps from different hierarchical levels of FPN along the channel direction and add a regression head at the end. However, simple concatenation lacks the covariance information among the concatenated features. Such information may help discern potential correlations indicating related features and could enhance the model's performance. Alternatively, feature maps from all hierarchical levels can be concatenated, and then SA applied to them, however, this is computationally expensive due to a large number of features. We address these problems by proposing an efficient feature fusion mechanism that combines the feature maps of consecutive hierarchies, going deeper into the network, and calculates SA among them. For that, it first concatenates the feature maps and then passes them through a 1×1 convolution layers to reduce the size of concatenated features. Next, it calculates SA on the resultant feature map along the concatenation direction i.e. channels. Mathematically, given F_i and F_{i-1}, the feature maps from the n and $n-1$ hierarchies of FPN, the EFFM mechanism can be formulated as:

$$\begin{aligned} F_C &= \text{Concat}(F_i, \text{AvgPool}(F_{i-1}))W_r \\ F_A &= \text{TSA}(F_C)W_s + F_C \\ F_O &= \text{FFN}(F_A)W_p + F_A \end{aligned}$$

where W_r is the 1×1 2D convolution layer, and F_C is the concatenated feature map. TSA is the 'Transposed Self-Attention' which is a modified version of conventional SA to calculate SA among channel direction, instead of spatial direction. It generates Q, K, and V embeddings of the shape $\mathbb{R}^{C \times HW}$, and then calculates self-attention using the formula i.e. $V\text{softmax}\left(\frac{KQ^T}{\alpha}\right)$. The attention map in TSA is of the shape $C \times C$ which has the covariance information among concatenated features. Like a conventional transformer block,

Table 1. Granular Level Data Distribution of Datasets. Note that for each section (anterior (A) and posterior(P)) of the vertebra (L1-L4), the distribution is strongly skewed towards the '0' score.

AAC Score	**DE / SE GE iDXA Dataset (1,916)**								**SE Hologic Dataset (508)**							
	L1		**L2**		**L3**		**L4**		**L1**		**L2**		**L3**		**L4**	
	A	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
0	1517	1498	1404	1360	1217	1120	1252	1013	484	461	453	439	401	383	365	322
1	291	245	355	328	319	339	290	315	22	39	34	50	70	91	89	118
2	64	89	100	121	186	210	151	250	2	7	21	17	23	27	43	57
3	44	84	57	107	194	247	223	338	0	1	0	2	14	7	11	11

the attention-added feature map F_A is then passed through the Feed Forward Network (FFN) to generate the output feature map F_O. W_s and W_p are the projection layers for TSA and FFN respectively. Experimental studies validate that this overall operation of EFFM improves performance.

3 Experiments and Results

We assess the effectiveness of our proposed framework using three sets of DXA VFA data: Dual Energy (DE) and Single Energy (SE) DXA variants from the GE iDXA machine, as well as SE DXA scans from the Hologic Horizon machine. The GE iDXA machine's SE and DE datasets comprise 1,916 VFA images with identical labels and AAC score distribution, however, the generated images originate from distinctly different methodologies used during the scan, resulting in varied distributions of pixel values. The Hologic Horizon's DXA dataset includes 508 VFA scans. Trained radiologists annotated the images using the Kauppila AAC-24 scoring method [11]. This method divides the abdominal aortic region in front of the L1 to L4 vertebrae into eight sections, delineating anterior and posterior sections for each vertebra (L1 to L4). Each section receives a score based on the extent of AAC. Specifically, a score of 0 is assigned if the AAC is less than one-third of the length of the adjacent vertebra, 1 if it exceeds one-third but falls short of two-thirds, and 3 if it extends from more than two-thirds to the full length of the adjacent vertebra. Thus, each section score ranges from 0 to 3. The distribution of AAC scores for each section (anterior or posterior) of all vertebrae L1-L4 is illustrated in Table 1, demonstrating a prominent skew towards 0 scores.

Implementation Details: Images in all three datasets are cropped from the top half to include the ROI only. For data augmentation, scaling, translation, rotation, and shear are used. 10-fold stratified cross-validation is used for performance analysis. EfficientNetV2S is used as CNN backbone in all experiments in this work, unless otherwise stated. The weighted Mean Square Error (MSE) loss function is used for regression of granular outputs with weight balancing to tackle the class imbalance problem. The loss function L_{Total} used is formulated as $\sum_{i=1}^{4} w_{Ai} L_{Ai} + \sum_{i=1}^{4} w_{Pi} L_{Pi}$, where L_{Ai} and L_{Pi} are the MSE losses for the anterior and posterior sections of the four vertebrae, i.e. L1 to L4, with respective balancing weights w_{Ai} and w_{Pi}. We implement the proposed model in Pytorch and train it on NVIDIA GeForce RTX 3080 Ti, using a batch size of 20, learning rate of 5e^{-4}, and Adam Optimizer.

Table 2. Comparative granular level analysis of the proposed framework with SOTA and Baseline models across three distinct DXA VFA datasets.

DE DXA Dataset GE iDXA (1,916 scans)					
Evaluation Metric	**Method**	**L1**	**L2**	**L3**	**L4**
Pearson Correlation ↑	Gilani et al. [15]	0.49	0.64	0.70	0.69
	Saleem et al. [16]	0.73	0.77	0.80	0.82
	Hybrid-FPN-AACNet	**0.77**	**0.80**	**0.83**	**0.84**
Kendall Tau ↑	Gilani et al. [15]	0.44	0.56	0.60	0.58
	Saleem et al. [16]	0.49	0.56	0.62	0.63
	Hybrid-FPN-AACNet	**0.51**	**0.58**	**0.65**	**0.66**
Mean Absolute Error ↓	Gilani et al. [15]	0.60	0.67	0.88	1.10
	Saleem et al. [16]	0.46	0.50	0.70	0.72
	Hybrid-FPN-AACNet	**0.43**	**0.47**	**0.69**	**0.70**
SE DXA Dataset GE iDXA (1,916 scans)					
Evaluation Metric	**Method**	**L1**	**L2**	**L3**	**L4**
Pearson Correlation ↑	Baseline	0.69	0.72	0.79	0.81
	Hybrid-FPN-AACNet	**0.79**	**0.81**	**0.84**	**0.85**
Kendall Tau ↑	Baseline	0.50	0.57	0.61	0.64
	Hybrid-FPN-AACNet	**0.53**	**0.60**	**0.67**	**0.68**
Mean Absolute Error ↓	Baseline	0.44	0.49	0.71	0.70
	Hybrid-FPN-AACNet	**0.40**	**0.45**	**0.67**	**0.67**
SE DXA Dataset Hologic Horizon (508 scans)					
Evaluation Metric	**Method**	**L1**	**L2**	**L3**	**L4**
Pearson Correlation ↑	Baseline	0.49	0.47	0.67	0.68
	Hybrid-FPN-AACNet	**0.60**	**0.59**	**0.74**	**0.70**
Kendall Tau ↑	Baseline	0.20	0.30	0.38	0.38
	Hybrid-FPN-AACNet	**0.27**	**0.37**	**0.42**	**0.40**
Mean Absolute Error ↓	Baseline	0.27	0.40	0.46	0.43
	Hybrid-FPN-AACNet	**0.21**	**0.35**	**0.42**	**0.40**

Comparison with Previous Works and Baseline: For the DE DXA VFA variant of the GE iDXA machine, we evaluate the effectiveness of our proposed framework with the previous works of Gilani et al. [15], and Saleem et al. [16] The model of Saleem et al. [16] is primarily designed for the single regression output indicating cumulative AAC score, however, we retrain the model after replacing its regression head with an eight outputs regression head (same as our framework). Using the same approach used by Gilani et al. [15], we compare the results for the combined output of the anterior and posterior sections of each vertebra. Table 2 illustrates a comparison between correlation and error metrics for ground truth at the granular level and predicted scores for all three methods. Our proposed model, Hybrid-FPN-AACNet, demonstrates notably superior performance across all four vertebrae L1, L2, L3, and L4 exhibiting increase in Pearson's correlation. For the SE DXA VFA variants of the GE iDXA machine, and the Hologic Horizon machine, we compare our approach with a baseline model only as no one else has previously reported results on them. For our baseline comparison, we employ the CNN-based FPN architecture (i.e. Hybrid-FPN-AACNet without DRSA Transformer and EFFM blocks) utilizing simple feature fusion and a regression head with 8 outputs. Table 2 presents a comparison of results, demonstrating that our proposed DRSA and EFFM modules improve

Table 3. Ablation Study - Experiment 1: Effect of using different attention mechanisms, and feature fusion techniques on performance. Experiment 2: Effect of feature map reduction factor r and window size p in DRSA on model performance.

Ablation Study Experiment 1								
Network Configuration	**L1A**	**L1P**	**L2A**	**L2P**	**L3A**	**L3P**	**L4A**	**L4P**
CNN-Backbone	0.54	0.61	0.56	0.67	0.70	0.70	0.69	0.72
CNN-Backbone FPN	0.55	0.61	0.59	0.65	0.71	0.69	0.71	0.73
CNN-Backbone FPN + HiLo Attention [27]	0.56	0.64	0.62	0.69	0.69	0.72	0.70	0.74
CNN-Backbone FPN + DRSA	0.60	0.66	0.63	0.72	0.71	0.75	0.71	0.77
CNN-Backbone FPN + DRSA + EFFM (Proposed)	**0.64**	**0.68**	**0.65**	**0.73**	**0.74**	**0.76**	**0.75**	**0.78**

Ablation Study Experiment 2												
FPN Hierarchy					**L1A**	**L1P**	**L2A**	**L2P**	**L3A**	**L3P**	**L4A**	**L4P**
Lvl1	**Lvl2**	**Lvl3**	**Lvl4**	**Lvl5**								
r=4, p=20	r=4, p=20	r=4, p=20	r=2, p=10	r=2, p=5	0.59	0.64	0.62	0.70	0.73	0.75	0.75	0.76
r=4, p=10	r=4, p=10	r=4, p=10	r=2, p=10	r=2, p=5	0.60	0.66	0.62	0.71	0.74	0.75	0.76	0.77
r=2, p=20	r=2, p=20	r=2, p=20	r=2, p=10	r=2, p=5	0.63	0.68	0.64	0.72	0.75	0.75	0.76	0.77
r=2, p=10	r=2, p=10	r=2, p=10	r=2, p=10	r=2, p=5	**0.64**	**0.68**	**0.65**	**0.73**	**0.75**	**0.76**	**0.76**	**0.78**

performance across all three datasets. Our proposed model has 22.02 million parameters while the CNN backbone without DRSA and EFFM has 19.84 million parameters. The inference time of our model increases by 91 ms (i.e. 408.08 ms), however, the average improvement in performance (PCC) is more than 10%. It is important to point out that our proposed model performs much better than the EfficientNetV2M backbone (without DRSA/ EFFM), which is more complex than ours (52.2 million parameters, 534.4 ms inference time).

Ablation Studies: For the ablation study, in the first experiment, we test different network configurations including the use of HiLo attention [27], DRSA, and EFFM. The Pearson Correlation Coefficient is used as an evaluation metric. Results in Table 3 show that the use of DRSA and EFFM improves the performance of the model, especially in the anterior and posterior sections of L1, and L2 vertebrae where calcification is usually in small amount and is difficult to differentiate from artifacts. In the second experiment, we test the effect of the feature map reduction factor (r) and the window size (p) in the DRSA block on model performance. For the DRSA blocks of the last two feature hierarchies, i.e. Lvl4 and Lvl5, we keep (r) and (p) fixed as feature maps in these levels already have small dimensions. The results in Table 3 show that the use of large reduction factor r deteriorates the performance of the model as it misses important information due to extensive downsampling. Optimal results are obtained with a window size of p of 10. A larger window size introduces more parameter complexity to the model. For further ablation study experiments, please refer to supplementary material.

4 Conclusion

This paper proposed a novel hybrid CNN-Transformer model Hybrid FPN-AACNet, based on the FPN hierarchy that employs a novel dual resolution self-attention block DRSA and efficient feature fusion module EFFM. The proposed model calculates and combines high frequency local context and low frequency broader context at each hierarchial level of FPN and effectively combines the feature maps using EFFM. The proposed architecture has achieved SOTA performance at granular level AAC detection.

Acknowledgments. De-identified labelled images were sourced for the ML (Project no: 03349 LEWIS) from several existing studies collecting VFAs. For the Hologic dataset, written informed consent was obtained from all participants. For Hologic images the HealthPartners Institutional Review Board [IRB] [#A20-149]) approved the use of the de-identified Hologic images or the ethics approval for the study was granted by the Edith Cowan University Human Research Ethics Committee (Project number: 20513 HODGSON). For the GE images the study was approved by the Health Research Ethics Board for the University of Manitoba (HREB H2004:017L, HS20121). The Manitoba Health Information Privacy Committee approved access to the iDXA GE data and waived the requirement for signed consent (HIPC 2016/2017–29). The study was supported by a National Health and Medical Research Council of Australia Ideas grant (APP1183570) and the Rady Innovation Fund, Rady Faculty of Health Sciences, University of Manitoba. The results and conclusions are those of the authors and no official endorsement by Manitoba Health and Seniors Care, or other data providers is intended or should be inferred. The salary of JRL is supported by a National Heart Foundation of Australia Future Leader Fellowship (ID: 102817). Also, SZG was partially funded by the Raine Priming Grant awarded by the Raine Medical Research Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. World Health Organization, "Cardiovascular diseases", `https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1`.
2. Strong, J.P., Malcom, G.T., McMahan, C.A., Tracy, R.E., Newman, W.P., Herderick, E.E., Cornhill, J.F.: Prevalence and Extent of Atherosclerosis in Adolescents and Young Adults: Implications for Prevention from the Pathobiological Determinants of Atherosclerosis in Youth Study. JAMA **281**(8), 727-735 (1999)
3. Golestani, R., Tio, RA., Zeebregts, C.J., Zeilstra, A., Dierckx, R.A., Boersma, H.H., Hillege, H.L., Slart, R.H.J.A.: Abdominal Aortic Calcification Detected by Dual X-ray Absorptiometry: A Strong Predictor for Cardiovascular Events. Annals of Medicine **42**(7), 539-545 (2010)
4. Lewis, J.R., Schousboe, J.T., Lim, W.H., Wong, G., Wilson, K.E., Zhu, K., Thompson, P.L., Kiel, D.P., Prince, R.L.: Long-Term Atherosclerotic Vascular Disease Risk and Prognosis in Elderly Women with Abdominal Aortic Calcification on Lateral Spine Images Captured During Bone Density Testing: A Prospective Study. Journal of Bone and Mineral Research **33**(6), 1001-1010 (2018)

5. Schousboe, J.T., Claflin, D., Barrett-Connor, E.: Association of Coronary Aortic Calcium with Abdominal Aortic Calcium Detected on Lateral Dual Energy X-ray Absorptiometry Spine Images. The American Journal of Cardiology **104**(3), 299-304 (2009)
6. Schousboe, J.T., Taylor, B.C., Kiel, D.P., Ensrud, K.E., Wilson, K.E., McCloskey, E.V.: Abdominal Aortic Calcification Detected on Lateral Spine Images from a Bone Densitometer Predicts Incident Myocardial Infarction or Stroke in Older Women. Journal of Bone and Mineral Research **23**(3), 409-416 (2008)
7. Leow, K., Szulc, P., Schousboe, J.T., Kiel, D.P., Pinto, A., Shaikh, H., Sawang, M., Sim, M., Bondonno, N., Hodgson, J.M., Sharma, A., Thompson, P.L., Prince, R.L., Craig, J.C., Lim, W.H., Wong, G., Lewis, J.R.: Prognostic Value of Abdominal Aortic Calcification: A Systematic Review and Meta-Analysis of Observational Studies. Journal of the American Heart Association **10**(2), e017205 (2021)
8. Isgum, I., Ginneken, B.V., Olree, M.: Automatic Detection of Calcifications in the Aorta from CT Scans of the Abdomen1: 3D Computer-Aided Diagnosis. Academic Radiology **11**(3), 247-257 (2004)
9. Setiawati, R., Chio, F.D., Rahardjo, P., Nasuto, M., Dimpudus, F.J., Guglielmi, G.: Quantitative Assessment of Abdominal Aortic Calcifications using Lateral Lumbar Radiograph, Dual-Energy X-ray Absorptiometry, and Quantitative Computed Tomography of the Spine. Journal of Clinical Densitometry **19**(2), 242-249 (2016)
10. Cecelja, M., Frost, M.L., Spector, T.D., Chowienczyk, P.: Abdominal Aortic Calcification Detection using Dual-Energy X-ray Absorptiometry: Validation Study in Healthy Women Compared to Computed Tomography. Calcified Tissue International **92**(6), 495-500 (2013)
11. Kauppila, L.I., Polak, J.F., Cupples, L.A., Hannan, M.T., Kiel, D.P., Wilson, P.W.F.: New Indices to Classify Location, Severity, and Progression of Calcific Lesions in the Abdominal Aorta: a 25-year Follow-up Study. Atherosclerosis, **132**(2), 245-250 (1997).
12. Chaplin, L., Cootes, T.: Automated scoring of Aortic Calcification in Vertebral Fracture Assessment Images. In: Medical Imaging 2019: Computer-Aided Diagnosis. vol. 10950, pp. 811–819. SPIE (2019)
13. Elmasri, K., Hicks, Y., Yang, X., Sun, X., Pettit, R., Evans, W.: Automatic Detection and Quantification of Abdominal Aortic Calcification in Dual Energy X-Ray Absorptiometry. Procedia Computer Science 96, 1011–1021 (2016)
14. Reid, S., Schousboe, J.T., Kimelman, D., Monchka, B.A., Jozani, M.J., Leslie, W.D.: Machine learning for automated abdominal aortic calcification scoring of DXA vertebral fracture assessment images: A pilot study. Bone 148, 115943 (2021)
15. Gilani, S.Z., Sharif, N., Suter, D., Schousboe, J.T., Reid, S., Leslie, W.D., Lewis, J.R.: Show, Attend and Detect: Towards Fine-Grained Assessment of Abdominal Aortic Calcification on Vertebral Fracture Assessment Scans. In: 2022 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 439-450 (2022).
16. Saleem, A., Ilyas, Z., Suter, D., Hassan, G.M., Reid, S., Schousboe, J.T., Prince, R., Leslie, W.D., Lewis, J.R., Gilani, S.Z.: SCOL: Supervised Contrastive Ordinal Loss for Abdominal Aortic Calcification Scoring on Vertebral Fracture Assessment Scans. In: 2023 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) pp. 273-283 (2023).
17. Sharif, N., Gilani, S.Z., Suter, D., Reid, S., Szulc, P., Kimelman, D., Monchka, B.A., Jozani, M.J., Hodgson, J.M., Sim, M., Zhu, K.: Machine learning for abdominal aortic calcification assessment from bone density machine-derived lateral spine images. EBioMedicine 94, (2023).

18. Dalla Via, J., Gebre, A.K., Smith, C., Gilani, Z., Suter, D., Sharif, N., Szulc, P., Schousboe, J.T., Kiel, D.P., Zhu, K., Leslie, W.D.: Machine-learning assessed abdominal aortic calcification is associated with long-term fall and fracture risk in community-dwelling older Australian women. Journal of Bone and Mineral Research, **38**(12), 1867-1876 (2023).
19. Westerhof, N., Lankhaar, J.W., Westerhof, B.E.: The arterial windkessel. Medical and Biological Engineering and Computing, **47**(2), 131-141 (2009).
20. Cho, I.J., Chang, H.J., Park, H.B., Heo, R., Shin, S., Shim, C.Y., Hong, G.R., Chung, N.: Aortic calcification is associated with arterial stiffening, left ventricular hypertrophy, and diastolic dysfunction in elderly male patients with hypertension. Journal of Hypertension, **33**(8), 1633-1641 (2015).
21. Cecelja, M., Jiang, B., Bevan, L., Frost, M.L., Spector, T.D., Chowienczyk, P.J.: Arterial stiffening relates to arterial calcification but not to noncalcified atheroma in women: a twin study. Journal of the American College of Cardiology, **57**(13), 1480-1486 (2011).
22. Mohiuddin, M.W., Laine, G.A., Quick, C.M.: Increase in pulse wavelength causes the systemic arterial tree to degenerate into a classical windkessel. American Journal of Physiology-Heart and Circulatory Physiology, **293**(2), H1164-H1171 (2007).
23. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117-2125 (2017).
24. Shareef, B., Xian, M., Vakanski, A., Wang, H.: Breast Ultrasound Tumor Classification Using a Hybrid Multitask CNN-Transformer Network. In: 2023 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 344-353 (2023).
25. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. In: 2023 IEEE Transactions on Medical Imaging (2023).
26. Wang, X., Ying, H., Xu, X., Cai, X., Zhang, M.: TransLiver: A Hybrid Transformer Model for Multi-phase Liver Lesion Classification. In: 2023 International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 329-338 (2023).
27. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with hilo attention. Advances in Neural Information Processing Systems, **35**, 14541-14554 (2022).