Gradient Guided Co-Retention Feature Pyramid Network for LDCT Image Denoising

Li Zhou[1][0000−0001−6989−8080], Dayang Wang[1], Yongshun Xu[1], Shuo Han[1], Bahareh Morovati[1], Shuyi Fan[1], and Hengyong Yu[1][0000−0002−5852−0813]

University of Massachusetts Lowell, Lowell MA 01854, USA
hengyong_yu@uml.edu

Abstract. Low-dose computed tomography (LDCT) reduces the risks of radiation exposure but introduces noise and artifacts into CT images. The Feature Pyramid Network (FPN) is a conventional method for extracting multi-scale feature maps from input images. While upper layers in FPN enhance semantic value, details become generalized with reduced spatial resolution at each layer. In this work, we propose a Gradient Guided Co-Retention Feature Pyramid Network (G2CR-FPN) to address the connection between spatial resolution and semantic value beyond feature maps extracted from LDCT images. The network is structured with three essential paths: the bottom-up path utilizes the FPN structure to generate the hierarchical feature maps, representing multi-scale spatial resolutions and semantic values. Meanwhile, the lateral path serves as a skip connection between feature maps with the same spatial resolution, while also functioning feature maps as directional gradients. This path incorporates a gradient approximation, deriving edge-like enhanced feature maps in horizontal and vertical directions. The top-down path incorporates a proposed co-retention block that learns the high-level semantic value embedded in the preceding map of the path. This learning process is guided by the directional gradient approximation of the high-resolution feature map from the bottom-up path. Experimental results on the clinical CT images demonstrated the promising performance of the model. Our code is available at: https://github.com/liz109/G2CR-FPN.

Keywords: LDCT denoising · Retention · Feature pyramid · Directional gradients.

1 Introduction

Compared with the normal-dose computed tomography (NDCT), low-dose computed tomography (LDCT) reduces the risks of ionizing radiation exposure but introduces noise and artifacts into the reconstructed images. A general solution is to develop denoising techniques to reduce or eliminate undesirable noise from CT images to improve their clarity and diagnostic values. Along this direction, deep learning models have been investigated for CT denoising, which learns an intrinsic feature map between noisy and clean CT images [7, 8, 28].

A pyramid structure [11], illustrated as the bottom-up path in Fig. 1 (A), stands out as a widely employed backbone for feature extraction. Compared with the columnar structure [2], where feature maps maintain the same size as the input image, the pyramid structure is cost-effective for addressing dense prediction tasks. This is because (1) the memory and computational costs are relatively low for the same input image size [24], and (2) feature maps are at different spatial scales and channel complexities, enabling the extraction of both coarse-grained and fine-grained information [15]. This versatility is advantageous for tasks at pixel level, where objects of interest may exhibit variations in size or scale within images while preserving consistent patterns, as often encountered in CT images. In leveraging feature maps learned from a neural network, many researchers have explored the utilization of the attention mechanism [22]. The self-attention operation depicts the intra-feature learning for a given input, while the guided-attention operation learns the inter-feature interactions across multiple inputs. Both of the operations, along with their combined form known as co-attention [13, 27, 29], exhibit competitive performance in modeling various computer vision tasks [2,12,21,23]. Notably, Sun *et al.* [19] recently redefined the attention mechanism as the retention mechanism in Natural Language Processing, showcasing competitive performance with the traditional attention mechanism and overcoming the quadratic computation complexity associated. As a followup work, in this paper we explore the possibility of the retention mechanism in pixel-level dense tasks, specifically focusing on inter- and intra-feature learning within the pyramid structure network. Targeting on model-based LDCT image denoising, previous studies often resulted in over-smoothed outputs with blurry edges. While edge enhancement methods [4, 5, 10] are introduced to preserve structural details in CT images, they compromise the diversity of feature maps and are confined to a fixed perceptual field, resembling columnar-like structures.

Thus, we draw inspiration from hierarchical feature maps, introducing the co-retention mechanism in two parts. First, we propose a self-retention operation for intra-feature learning, focusing on the integral perceptual field of a feature map itself. This operation captures dependencies within the feature map, enhancing its interpretability. Meanwhile, we introduce a guided-retention operation for inter-feature learning, emphasizing mutual perceptual fields between two feature maps. This operation uncovers interdependencies between a feature map with high-level semantic value and a feature map with high resolution. Specifically, we employ a directional gradient approximation method, similar to the edge detection method [18], on the high-resolution feature map. This process generates edge-like enhanced feature maps for adaptive guided feature learning, wherein the method decomposes the high-resolution features into horizontal and vertical directions. As a result, our proposed model, the Gradient Guided Co-Retention Feature Pyramid Network (G2CR-FPN), effectively addresses the LDCT denoising through comprehensive feature learning and detail preservation. The key contributions are summarized as follows: 1) We propose a G2CR-FPN, attempting to generate multi-scale feature maps and learn intrinsic information

by bridging feature maps with high-level semantic value and high resolution for LDCT denoising. The model exhibits promising performance in the experimental results. 2) We introduce a co-retention mechanism for pixel-level dense tasks, comprising self-retention and guided-retention operations. The new mechanism focuses on intra- and inter-feature learning within hierarchical feature maps. 3) We validate the effectiveness of directional gradient approximation in feature maps. The introduced edge-like directions enhance structures within the feature maps, mitigating over-smoothing issues.

2 Methods

2.1 Overall Structure

Our objective is to introduce feature learning techniques for the pyramid structure of feature maps, enhancing interpretability and preserving details for pixel-level denoising tasks. The overall structure of G2CR-FPN is illustrated in Fig. 1 (A), and a typical level of G2CR-FPN, depicting the intra- and inter-feature learning among feature maps, is shown in Fig. 1 (B).

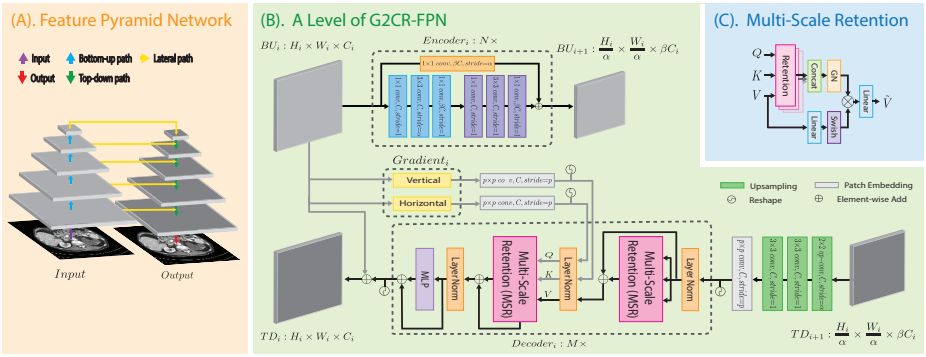

Fig. 1. Overview of the G2CR-FPN. (A) the overall structure and paths of the model. (B) the i-th level of G2CR-FPN, where connections between feature maps labeled as BU represent the bottom-up path, connections between TD represent the top-down path, and connections between BU and TD represent the lateral path. (C) multi-scale retention within a Decoder.

The proposed structure comprises five paths. To begin with, an image of size $H \times W \times 1$ is processed into a feature map $BU_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ through an input path. The path consists of two convolutional layers, each is connected to batch normalization and ReLU activation. Moving through the bottom-up path, a residual encoder is introduced to control the scales of feature maps, resulting in more generalized details. Meanwhile, the lateral path acts as a skip connection between feature maps with the same spatial resolution, while also

emphasizing directional gradient approximation represented as horizontal and vertical directions. Advancing through the top-down path, a co-retention decoder is introduced to incorporate high-level semantic value. This learning process is guided by directional gradient approximation from the high-resolution feature map. In the end, an output path integrates features into a denoised image of size $H \times W \times 1$. The output path consists of operations including layer normalization, a convolutional layer, batch normalization, and another convolutional layer.

2.2 Residual Encoder

The bottom-up path involves feed-forward convolutional computations for feature extraction, generating a pyramid of feature maps at various scales using a spatial shrinking factor (α) and a channel expanding factor (β). Inspired by the ResNets [6], a residual encoder is employed to learn input features. The encoder includes a residual connection and sequences of convolutional layers, batch normalization, and ReLU activation (see Fig. 1 (B)). Specifically, the second convolutional layer in the block reduces spatial dimensions using a stride of α, while the last layer and residual connection expand channels via β. We define a stack of encoders as a level in the pyramid structure where the output sizes of the encoders are consistent, and opt for the last feature map at each level. We denote an encoder layer as $Encoder_i$, with the input feature map labeled $BU_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ and the output feature map as $BU_{i+1} \in \mathbb{R}^{(H_i/\alpha) \times (W_i/\alpha) \times (\beta C_i)}$.

2.3 Directional Gradient Approximation

The purpose of the gradient approximation in the lateral path is to emphasize feature semantics with high spatial frequency and enrich the expression of feature maps in multiple view directions. In this work, we introduce a modification of the Sobel edge detection operator [18], which computes image intensity gradients through isotropic 3x3 kernels. The kernels now work with learnable factors, enabling adaptive optimization during the training process to generate edge-like feature maps. In the directional gradient operation $Gradient_i$, gradient approximations are computed in horizontal and vertical kernels, respectively. The gradient approximations are expressed as

$$BU_i^{Hor} = (w_i^{Hor} \cdot \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}) * BU_i, \quad BU_i^{Ver} = (w_i^{Ver} \cdot \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}) * BU_i, \quad (1)$$

where BU_i^{Hor} and $BU_i^{Ver} \in \mathbb{R}^{H_i \times W_i \times C_i}$ respectively denote the horizontal and vertical gradient approximations, w_i^{Hor} and $w_i^{Ver} \in \mathbb{R}$ are learnable factors, and $*$ denotes the convolution operation.

2.4 Co-Retention Decoder

The retention mechanism is a pivotal module in the Retentive Network [19], encoding sequences in an autoregressive manner and exhibiting dual forms of

recurrence and parallelism. Given an input sequence $X = [x_1, ..., x_{|x|}] \in \mathbb{R}^{|x| \times d}$ with an embedding dimension d, we formulate a sequence-to-sequence mapping $f : X_n \mapsto O_n$ along with linear representations of value (V_n), query (Q_n) and key (K_n) through state S_n. The linear representations are formulated as

$$V = XW^V,\ Q = (XW^Q) \odot \Theta,\ K = (XW^K) \odot \overline{\Theta}, \tag{2}$$

where $W^V, W^Q, W^K \in \mathbb{R}^{d \times d}$ are learnable weighting matrices, Θ is Extrapolatable Position Embedding (xPos) proposed by Sun *et al.* [20], $\overline{\Theta}$ is the conjugate of Θ, and $\odot$ is the element-wise multiplication.

Considering the parallel manner, the retention mechanism can be written as

$$Retention(X) = GN((QK^\intercal \odot D)V), \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}, \tag{3}$$

where GN is short for Group Normalization [26], $D \in \mathbb{R}^{|x| \times |x|}$ denotes a decay mask, and γ is a scalar.

Instead of employing a single parallel retention mechanism to obtain representations as value, query, and key, characterized by parameter matrices W^V, W^Q and W^K, it is advantageous to project the representations H times in each layer, each time using different parameter matrices W_h^V, W_h^Q and $W_h^K \in \mathbb{R}^{d_{head} \times d_{head}}$, where $d_{head} = d/H$. In addition, the retention is expanded into multi-scale retention (MSR), involving H retention heads operating in parallel with the scalars $\Gamma = \left\{\gamma_h\right\}_{h=1}^{H}$, respectively. Then, the outputs are concatenated and normalized. The MSR layer is defined as

$$\begin{aligned} &Head_h = Retention(X, \gamma_h), \\ &Y = GN(Concat(Head_1, ..., Head_H)), \\ &MSR(X) = (Swish(XW^G) \odot Y)W^O, \end{aligned} \tag{4}$$

where W^G and $W^O \in \mathbb{R}^{d \times d}$ are learnable matrices, GN normalizes each head separately, and $Swish$ [16] activation improves the non-linearity of layers. To incorporate the co-retention mechanism, which is designed to receive different sequences of token embeddings as input, we have to redefine the function as

$$\begin{aligned} &Head_h = Retention(V, Q, K, \gamma_h), \\ &MSR(V, Q, K) = (Swish(VW^G) \odot Y)W^O, \end{aligned} \tag{5}$$

where Y follows the same operations as shown in Equation (4).

In the following, we introduce the components of a co-retention decoder $Decoder_i$, as shown in Fig. 1. The decoder comprises self-retention and guided-retention operations. Operations are alternatively connected, accompanied by layer normalization (LN) and a residual connection. For an input feature map $TD_{i+1} \in \mathbb{R}^{(H_i/\alpha) \times (W_i/\alpha) \times (\beta C_i)}$, we use an upsampling layer followed by a patch embedding operation before each decoder block. Inspired by the ViT [2], the patch embedding operation involves partitioning the input into 2D patches and

flattening/reshaping them into the patch embeddings $\tilde{TD}_{i+1} \in \mathbb{R}^{K_i^2 \times P^2 \times C_i}$. Specifically, we partition an input into K_i^2 evenly spaced patches by a fixed patch size of $P \times P$ in different levels, enabling inter-feature learning within matched perceptual fields among feature maps. The input is partitioned using a kernel size $K_i = H_i/P = W_i/P$ in a convolutional layer. The embedding operation is also employed in the directional gradient approximations, resulting $\tilde{BU}_i^{Hor}$ and $\tilde{BU}_i^{Ver}$. The embedded horizontal and vertical approximations respectively serve as Q and K in the second MSR of the decoder.

3 Experiments and Results

3.1 Experimental Setup

Datasets: We conduct experiments on the dataset from the 2016 NIH-AAPM-Mayo Clinic LDCT Grand Challenge [14], consisting of 2,378 CT images with a slice thickness of 3.0 mm. The dataset was collected from ten different patients. We randomly select subject 'L506' for testing, and images from the remaining subjects are for training.

Implementation details: In the G2CR-FPN model, there are five levels, denoted as $L = 5$. At each level, the structure includes a stack of two encoders ($N = 2$) in the BU path and a single decoder ($M = 1$) with $H = 8$ heads in the TD path. The spatial factor (α) and the channel factor (β) are both set to 2. The dimension of each patch is set as $P = 32$. The model is trained with MSE loss function and Adam optimizer with default settings for at most 200 epochs, and the best model with the minimal loss is saved. The learning rate is initially set as 0.001 and is halved for every 3,000 steps in the training stage.

Evaluation metrics: For quantitative assessments, we employ two conventional metrics: root mean square error (RMSE) and structural similarity (SSIM) [25]. Additionally, we introduce the Edge Structural Similarity Index (E-SSIM) to evaluate the performance of the gradient approximation operation and the guided-retention operation. E-SSIM combines the SSIMs for both edge and non-edge regions, where $E\text{-}SSIM = 0.5 \times edge\text{-}SSIM + 0.5 \times non\text{-}edge\text{-}SSIM$. The edges are detected by the Sobel filter. The average results for the testing subject are presented within the Hounsfield Unit (HU) window of $[-160, 240]$.

3.2 Experimental Results

To evaluate the performance of the proposed G2CR-FPN model, we compare it against several state-of-the-art models: REDCNN [1], MAPNN [17], SwinIR [9], and SUNet [3]. The parameters for the competing methods are configured according to the guidelines disclosed in the corresponding papers. Table 1 summarizes the quantitative results, showing the mean±SDs (standard deviations) from the testing images using these methods. The LDCT images initially had the lowest scores due to the degradation from low dose radiation. All methods improved the scores, with G2CR-FPN achieving the best scores.

Table 1. Comparison with the SOTA and Ablation Study (mean±SDs). **Bold**:Best.

Method	RMSE ↓	SSIM ↑	E-SSIM ↑	edge-SSIM ↑
LDCT	10.4833 ± 1.5358	0.9359 ± 0.0170	0.9164 ± 0.0322	0.9216 ± 0.0334
REDCNN	9.0390 ± 2.1339	0.9466 ± 0.0224	0.9257 ± 0.0283	0.9287 ± 0.0300
MAPNN	8.1332 ± 1.6802	0.9528 ± 0.0188	0.9303 ± 0.0256	0.9319 ± 0.0267
SwinIR	7.5886 ± 1.6717	0.9556 ± 0.0182	0.9317 ± 0.0248	0.9341 ± 0.0259
SUNet	7.3576 ± 1.5899	0.9584 ± 0.0175	0.9331 ± 0.0250	0.9376 ± 0.0257
CA-FPN	7.7318 ± 1.6450	0.9552 ± 0.0179	0.9301 ± 0.0249	0.9328 ± 0.0258
G2CA-FPN	7.4341 ± 1.6191	0.9559 ± 0.0186	0.9346 ± 0.0254	0.9362 ± 0.0263
CR-FPN	7.4127 ± 1.5902	0.9585 ± 0.0172	0.9340 ± 0.0242	0.9373 ± 0.0251
G2CR-FPN	$\mathbf{7.0516 \pm 1.5358}$	$\mathbf{0.9602 \pm 0.0170}$	$\mathbf{0.9400 \pm 0.0253}$	$\mathbf{0.9420 \pm 0.0258}$

Moreover, a representative slice with lesions from Case L506 is selected to evaluate the performance of the aforementioned models. As shown in Fig. 2, the first row displays the denoised images from each model, with the red box indicating the region of interest (ROI) zoomed in the second row. It can be clearly observed that all the methods suppress image noise to various degrees. The SUNet and G2CR-FPN generate clearer noise-free images, and they can better discriminate low contrast regions in soft tissues than other models. To further differentiate the performance of the models, we conduct noise power spectrum (NPS) analysis on Fig. 2. As depicted in Fig. 3 (A), all images exhibit similar power spectra at low spatial frequencies, while differences became visible as frequencies increase. The G2CR-FPN achieves the lowest noise in the full frequency range. In summary, our proposed G2CR-FPN model is competitive with the state-of-the-art methods in LDCT denoising, evidenced by both quantitative metrics and qualitative visual results.

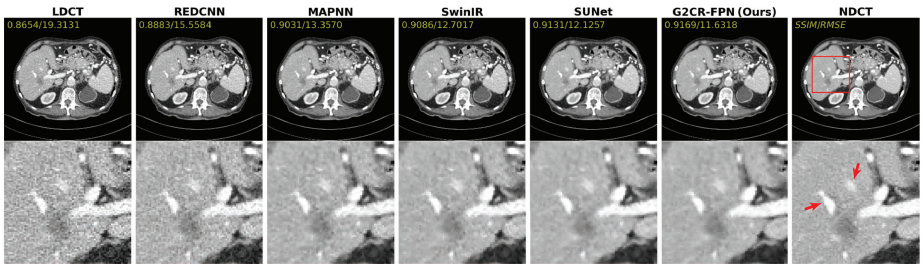

Fig. 2. Qualitative comparison from Case L506. The red boxes indicate the zoomed ROIs and the arrows point to the target lesions. The display window is [-160,240] HU.

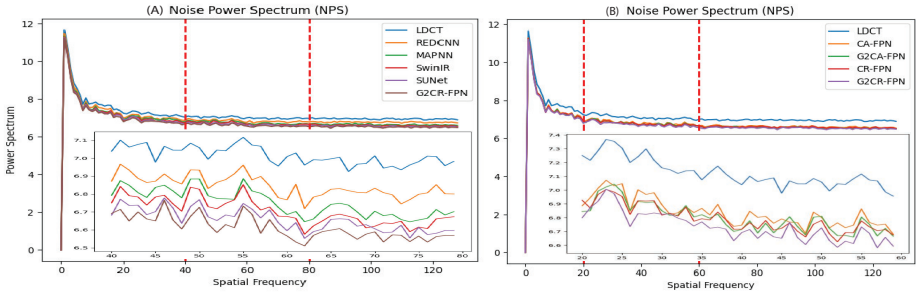

Fig. 3. Noise power spectrum analysis of the representative slice for (A) the comparison with the SOTA methods, and (B) the ablation study.

3.3 Ablation Study

As the aforementioned, the retention network is a successor to the transformer architecture. Our study aims to demonstrate the effectiveness of the Co-Retention module within the G2CR-FPN model by comparing it to the attention mechanism [22]. Specifically, we evaluate the Co-Retention decoder (G2CR-FPN) against the Co-Attention decoder (G2CA-FPN) within the architecture shown in Fig.1 (B). Additionally, we aim to validate the impact of the trainable directional gradient approximation in the lateral path. To do so, we compare with models without gradient operations while maintaining consistent remaining structures, labeled as CR-FPN and CA-FPN respectively.

The quantitative results are also summarized in Table 1. The Co-Retention-based models (G2CR-FPN and CR-FPN) outperform the Co-Attention-based models (G2CA-FPN and CA-FPN) in terms of both SSIM and RMSE. Furthermore, the inclusion of gradient operations leads to quantitative improvements in both G2CA-FPN and G2CR-FPN models. Although the improvement between G2CR-FPN and CR-FPN is small in terms of SSIM (0.0017), the RMSE is reduced by 5%, demonstrating that the combined mechanism is better and therefore the best model for LDCT denoising. A detailed examination of the outputs is provided in Fig. 3 (B) in terms of NPS analysis. All the images denoised using our methods exhibit significantly lower power spectra than the LDCT image.

4 Conclusions

We introduce a novel gradient guided co-retention feature pyramid network (G2CR-FPN) for LDCT image denoising and demonstrate how the proposed directional feature gradient approximation and co-retention mechanisms cooperatively learn feature maps in high resolution and high semantic value. Specifically, we show how the gradient approximation operation perceives a feature map in horizontal and vertical directions, and how the co-retention mechanism can tackle high inter- and intra- feature interactions among different scales of

levels in the pyramid structure network. Our experimental results indicate the potential of these mechanisms and achieve encouraging results for LDCT denoising.

Acknowledgments. This work was supported in part by NIH/NIBIB under grants R01EB032807 and R01EB034737, and NIH/NCI under grant R21CA264772.

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

References

1. Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G.: Low-dose ct denoising with convolutional neural network. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 143–146. IEEE (2017)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Fan, C.M., Liu, T.J., Liu, K.H.: Sunet: swin transformer unet for image denoising. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 2333–2337. IEEE (2022)
4. Gholizadeh-Ansari, M., Alirezaie, J., Babyn, P.: Deep learning for low-dose ct denoising using perceptual loss and edge detection layer. Journal of digital imaging **33**, 504–515 (2020)
5. Han, S., Zhao, Y., Li, F., Ji, D., Li, Y., Zheng, M., Lv, W., Xin, X., Zhao, X., Qi, B., et al.: Dual-path deep learning reconstruction framework for propagation-based x-ray phase–contrast computed tomography with sparse-view projections. Optics Letters **46**(15), 3552–3555 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Immonen, E., Wong, J., Nieminen, M., Kekkonen, L., Roine, S., Törnroos, S., Lanca, L., Guan, F., Metsälä, E.: The use of deep learning towards dose optimization in low-dose computed tomography: A scoping review. Radiography **28**(1), 208–214 (2022)
8. Kulathilake, K.S.H., Abdullah, N.A., Sabri, A.Q.M., Lai, K.W.: A review on deep learning approaches for low-dose computed tomography restoration. Complex & Intelligent Systems **9**(3), 2713–2745 (2023)
9. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
10. Liang, T., Jin, Y., Li, Y., Wang, T.: Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising. In: 2020 15th IEEE International Conference on Signal Processing (ICSP). vol. 1, pp. 193–198. IEEE (2020)
11. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

12. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
13. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems **29** (2016)
14. McCollough, C.H., Bartley, A.C., Carter, R.E., Chen, B., Drees, T.A., Edwards, P., Holmes III, D.R., Huang, A.E., Khan, F., Leng, S., et al.: Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. Medical physics **44**(10), e339–e352 (2017)
15. Morovati, B., Lashgari, R., Hajihasani, M., Shabani, H.: Reduced deep convolutional activation features (r-decaf) in histopathology images to improve the classification performance for breast cancer diagnosis. arXiv preprint arXiv:2301.01931 (2023)
16. Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 **7**(1), 5 (2017)
17. Shan, H., Padole, A., Homayounieh, F., Kruger, U., Khera, R.D., Nitiwarangkul, C., Kalra, M.K., Wang, G.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose ct image reconstruction. Nature Machine Intelligence **1**(6), 269–276 (2019)
18. Sobel, I.: An isotropic 3× 3 image gradient operater. Machine vision for three-dimensional scenes pp. 376–379 (1990)
19. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., Wei, F.: Retentive network: A successor to transformer for large language models. arXiv preprint arXiv:2307.08621 (2023)
20. Sun, Y., Dong, L., Patra, B., Ma, S., Huang, S., Benhaim, A., Chaudhary, V., Song, X., Wei, F.: A length-extrapolatable transformer. arXiv preprint arXiv:2212.10554 (2022)
21. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
22. 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)
23. Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., Yu, H.: Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising. Physics in Medicine & Biology **68**(6), 065012 (2023)
24. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
26. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
27. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6281–6290 (2019)

28. Zhang, F., Liu, J., Liu, Y., Zhang, X.: Research progress of deep learning in low-dose ct image denoising. Radiation Protection Dosimetry **199**(4), 337–346 (2023)
29. Zhou, L., Luo, Y.: Deep features fusion with mutual attention transformer for skin lesion diagnosis. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3797–3801. IEEE (2021)