EasyChair Preprint

№ 6259

Perception-Aware Losses Facilitate CT Denoising and Artifact Removal

Suhita Ghosh, Andreas Krug, Georg Rose and Sebastian Stober

August 7, 2021

Perception-Aware Losses Facilitate CT Denoising and Artifact Removal

Suhita Ghosh*†, Andreas Krug*, Georg Rose†, and Sebastian Stober*
*AILab, Faculty of Computer Science, Otto-von-Guericke University, Magdeburg, Germany
†Institute for Medical Engineering and Research Campus STIMULATE, Otto-von-Guericke University, Magdeburg, Germany
{suhita.ghosh, andreas.krug, georg.rose, stober}@ovgu.de

***Abstract*—The concerns over radiation-related health risks associated with the increasing use of computed tomography (CT) have accelerated the development of low-dose strategies. There is a higher need for low dosage in interventional applications as repeated scanning is performed. However, using the noisier and undersampled low-dose datasets, the standard reconstruction algorithms produce low-resolution images with severe streaking artifacts. This adversely affects the CT assisted interventions. Recently, variational autoencoders (VAEs) have achieved state-of-the-art results for the reconstruction of high fidelity images. The existing VAE approaches typically use mean squared error (MSE) as the loss, because it is convex and differentiable. However, pixelwise MSE does not capture the perceptual quality difference between the target and model predictions. In this work, we propose two simple but effective MSE based perception-aware losses, which facilitate a better reconstruction quality. The proposed losses are motivated by perceptual fidelity measures used in image quality assessment. One of the losses involves calculation of the MSE in the spectral domain. The other involves calculation of the MSE in the pixel space and the Laplacian of Gaussian transformed domain. We use a hierarchical vector-quantized VAE equipped with the perception-aware losses for the artifact removal task. The best performing perception-aware loss improves the structural similarity index measure (SSIM) from 0.74 to 0.80. Further, we provide an analysis of the role of the pertinent components of the architecture in the denoising and artifact removal task.**

***Index Terms*—computed tomography, perception-aware, image reconstruction, deep learning, denoising, artifact removal**

I. Introduction

Computed tomography (CT) is the most frequently used tomographic method in many countries due to its wide availability, easy usage and minimal contraindications [1]. It plays an important role in providing assistance for various interventional procedures, including biopsy, tumour ablation, catheter placements and orthopaedic surgeries [2]. However, the burgeoning use of CT increases the radiation exposure related health risks and may induce cancer, especially for the paediatric patients [3]. These growing concerns acted as the driving force for the introduction of various dose reduction strategies. The most common dose reduction technique is lowering of the tube current or voltage. However, this strategy leads to noisier measurements (projections) with decreased signal-to-noise ratio due to increased electronic readout noise. Another potential dose reduction strategy is compressed-sensing motivated sparse sampling (SS) [4], where a reduced number of projections are acquired while maintaining the routine dose intensity. Although not clinically introduced, SS has shown potential to provide robust bone-mineral deposits [5] and facilitate faster data collection. The combination of lower tube current and sparse sampled projections will deliver ultra low-dose, desired for real-time CT assisted interventions.

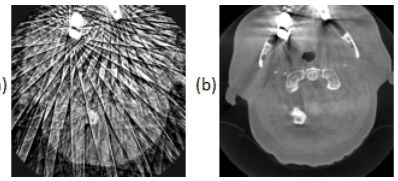

Fig. 1. A 2D slice of FDK reconstructed head CT volume (a) under sparse sampling and low beam intensity (b) using all routine dose projections. Both of them suffer from metal artifacts.

For decades, the conventional analytical methods such as filtered back projection (FBP) and Feldkamp (FDK) [6] are being used in clinics, as they need less computational power and time. However, the reconstructions using low-dose strategies and analytical methods suffer from severe streaking artifacts, as shown in Fig. I. This makes the reconstructions clinically useless and may hinder interventional guidance. To overcome this problem, different variants of iterative methods [4] were proposed, which can integrate prior knowledge, such as diagnostic quality pre-operative CT examinations or information about metallic implants. However, the iterative methods are much slower and more computational intensive than FDK, as they require to repeat the projection and back-projection operations multiple times. Recently, many deep learning (DL) methods have been proposed for the CT denoising task [7]. DL methods can incorporate prior knowledge, such as the FBP/FDK reconstructed image. Also, DL methods used as a post-processing tool takes fraction of a second for the inference, desirable for real-time interventions.

In this work, we propose a hierarchical vector quantized VAE (VQ-VAE) [8] based architecture to generate high quality images from the corrupted CT images acquired under low-dose and sparse sampling protocols. VQ-VAE is a DL based generative model which has achieved state-of-the-art results for various computer vision tasks [9]. VQ-VAE produces promising results, but the choice of loss function plays an important role for the specific task. Therefore, we introduce two perception-aware MSE based losses inspired by their usage in the image quality assessment tasks. The framework equipped with the proposed losses reconstructs high fidelity human head CT images, where some of them have metal fillings, making the denoising and artifact removal problem particularly challenging. Lastly, we demonstrate the role of each pertinent component of the hierarchy in the artifact-free reconstruction task.

II. Related Work

A. CT Reconstruction Using DL Approaches

The review on DL methods for CT reconstruction [7] portrays that significant efforts have been made to fuse the conventional

reconstruction algorithms with DL. The proposed DL based approaches can be divided into three categories: (i) unrolled iteration, (ii) domain-transform, and (iii) post-processing. The algorithms pertaining to the first category use DL models to learn some components of the iterative algorithms. The first DL method to unroll the optimization algorithm is ADMM-Net [10], where the tuning parameters and linear operator are learned from the training data. Some other approaches learn the regularizer [11] or model the projector which projects onto a set of feasible images, performed iteratively [12]. These methods produce diagnostic-quality images, but are slow for fast-imaging protocols owing to the inherent method's iterative nature.

The domain-transform approach estimates a direct mapping from projections to image domain. The method requires a lot of training data due to the huge difference between the projections and image space distributions. Zhu et al. proposed one such method AUTOMAP [13], which uses two fully-connected layers as the initial layers. The first layer learns a mapping between the sinograms and pixel domain. This approach is not feasible to reconstruct 3D volumes even using sparse sampling, considering its gigantic computational requirement [14]. Few other domain-transform approaches were proposed, where one X-ray image [15] or two bi-planar projections [16] were used to produce 3D CT volumes. However, both of them used clean projections with no metal artifacts. Further, the bi-planar network was trained with knee bone segmentation, which is difficult to acquire for every patient.

Most of the DL approaches for CT reconstruction belong to the post-processing category, which are primarily based on residual auto-encoder architectures [17, 18, 19]. In post-processing, a mapping is estimated between the low quality image/projection and the high quality image/projection. Few convolutional neural network (CNN) approaches performed denoising in the image domain, where a network was trained to produce cleaner images from the corrupted FBP reconstructions [19, 20, 21]. These networks were either trained on 3D volume patches or 2D slices. Some other works applied denoising on the projections prior to reconstruction [19, 22]. Lee et al. proposed a hybrid denoising framework [22], where the pooling layers were replaced with wavelet transforms and two residual networks were trained to perform denoising in both projections and image domain. Yang et al. proposed a Generative Adversarial Network (GAN) based approach [23] to denoise the low-dose scans. The GAN approach used a loss function as the combination of Wasserstein distance and the difference between a pretrained VGG network [24] outputs for noisier input image and target.

B. Applications of Discrete Cosine Transform (DCT)

DCT is a linear transform which decomposes an image into its spatial spectral components. It is widely used for data compression due to its energy compaction capabilities. Recently, it has been used in various tasks such as, classification accuracy improvement [25], face recognition [26], deep-fake identification [27], model compression [28, 29], faster training and convergence [30], produce harmonic convolutional blocks and reduce overfitting [29]. Giudice et al. [27] detected the GAN generated images by modelling the non-zero frequency coefficients of DCT as a zero-centred Laplacian distribution. The β statistics of the distribution was exploited to identify the GAN-specific spurious frequency, absent in normal images. Networks trained on 8×8 blockwise DCT coefficients taken directly from the JPEG images demonstrated significantly faster convergence and better accuracy for most of the cases [31]. DCT-based similarity metrics are used to assess the perceptual quality of the images [32] and video [33], where the distortion is measured by a $l2$-norm in the DCT domain.

III. Methodology

A. VQ-VAE Framework

VAE is a fusion of autoencoder and variational Bayes, where the input **x**, target **y** and embedding **e** are random variables. VAEs approximate the underlying data distribution instead of embedding all information explicitly, which makes them require fewer examples and less prone to overfitting than discriminative models. This makes them particularly suitable for small datasets. The VAE models produce blurry reconstructions as the encoder cannot precisely distinguish between multiple training samples, which causes the decoder weights to be spread across for the same samples [34]. VAEs face another problem of 'posterior collapse' due to the discrepancy between the posterior and prior distribution. This leads the decoder to utilize the information only from a subset of latent dimensions [35].

VQ-VAE uses discrete embeddings instead of a continuous distribution, which prevents the posterior collapse and stabilizes training [8]. The discrete embedding space is achieved through vector quantization, where K prototype vectors are indexed in an embedding space $\mathbf{e} \in R^{K \times D}$, where D is the dimension of each vector. The encoder E(.) provides a non-linear mapping between the input space (**x**) and a vector E(**x**), which is quantized using **e**. The optimal prototype vector for the encoder output E(**x**) is found by nearest-neighbour search between the E(**x**) and the prototype vectors in **e**. The decoder produces the reconstruction using the prototype vector from **e** through another non-linear function.

B. Proposed Hierarchical VQ-VAE Framework

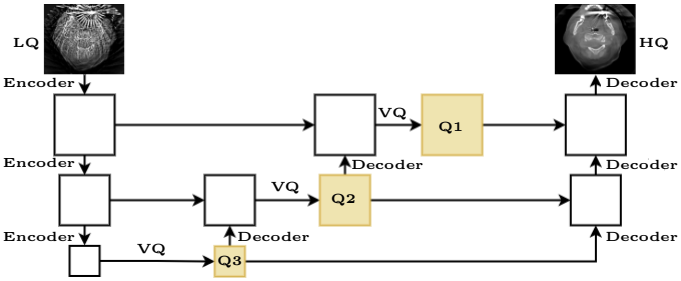

Fig. 2. A simple representation of the proposed hierarchical VQ-VAE concept

In our denoising and artifact removal task, the VQ-VAE network learns a transformation of the low quality (LQ) image to a higher quality (HQ) image. We modified VQ-VAE to have a hierarchy of three levels of quantized embeddings (Q1, Q2, Q3), as shown in Fig. 2, where VQ denotes vector quantization. The upper embeddings are conditioned on the lower ones as done in [9]. The hierarchical concept facilitates each embedding space to capture non-redundant representations which act complementary to each other. In this way, we intend to encourage each embedding level to capture separate representations. We hypothesize that the upper embedding space (Q1) models artifact removal, the

middle one (Q2) captures the local information such as edges and texture and the bottom one (Q3) captures the global structure. The detailed architecture is described in Appendix B.

C. Loss Formulation

The VQ-VAE [9] objective has multiple loss components, as shown in Equation 1. The first component is the reconstruction loss, which is the $l2$-norm between the model prediction ($\hat{\mathbf{y}}$) and the target ($\mathbf{y}$), optimising both encoder and decoder. For clarity, HQ is denoted by $\mathbf{y}$ in all equations. The second component is the commitment loss which affects only the encoder weights. It prevents frequent reassignment of a prototype vector to the encoder output, which prevents the explosion of the embedding space. For our task, there exists one commitment loss for each embedding level $l \in L$, as shown in Equation 1. Further, β denotes the change-reluctance hyperparameter [9], $\mathbf{e}_l$ is the embedding at level l, $E_l(\mathbf{x}_l)$ is the encoder output at level l using the input $\mathbf{x}_l$, sg is stopgradient operator [8]. The embedding vectors at all levels are learned through exponential moving average for faster convergence [9].

$$Loss = ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 + \beta \sum_{l=1}^{L} ||sg[\mathbf{e}_l] - E_l(\mathbf{x}_l)||_2^2 \quad (1)$$

The VAE and other models designed for reconstruction task typically use the classical MSE to measure the fidelity of the pixels, as it has desirable mathematical properties and is easily computed. However, pixelwise MSE does not capture the structural relationship in a pixel-neighbourhood, dissimilar to how humans perceive an image [36]. Therefore, MSE is considered as an unreliable metric in image quality assessment studies. On the other hand, the accepted perceptual metrics have at least one undesirable property when being used as objective function: non-convex, non-differentiable, invalid distance metrics, or have complicated gradient computation [37]. Therefore, in this work we propose two MSE based perception-aware losses. The losses enjoy the properties of MSE and compared to MSE, they are more aligned with how humans perceive the difference between two images.

1) Laplacian of Gaussian (LoG) MSE Loss: LoG is widely used for capturing sudden intensity changes (edge detection). It is calculated as the second derivative of an image after application of a Gaussian filter to the image, to reduce false edge detection. We propose a loss LoG MSE (Loss_{LoG}) as a combination of pixelwise MSE and $l2$-norm calculated in the LoG transformed pixel space, as shown in Equation 2, where λ is a hyperparameter for tuning LoG's contribution. The loss penalizes the disagreement between the zero crossings of LoG transformed target and the model predictions. This makes it more sensitive to the edges or structures wrongly reconstructed by the model, making it more perceptually aware than a pixelwise MSE.

$$Loss_{LoG}(\mathbf{y}, \hat{\mathbf{y}}) = MSE(\mathbf{y}, \hat{\mathbf{y}}) + \lambda\ MSE(LoG(\mathbf{y}), LoG(\hat{\mathbf{y}})) \quad (2)$$

2) DCT MSE Loss: DCT has been used in image quality assessment tasks [33], which motivated us to propose a loss where $l2$-norm is calculated in the spectral domain, as shown in Equation 3. Karhunen–Loève transform (KLT) is considered the optimal transform to extract weak signals hidden in any type of noise, where even Fast Fourier Transform does not work [38]. DCT closely approximates KLT under Markovian conditions [39], which is generally true for any CT image. DCT has the additional advantage of fixed basis images, unlike KLT. The components of the DCT spectrum have varying importance (amplitudes), based on their contribution to the visual quality of the image. Therefore, a huge disagreement between the model prediction and target for the important DCT coefficients will result in higher penalty. This facilitates the model to focus on the pertinent areas of the image.

$$Loss_{dct}(\mathbf{y}, \hat{\mathbf{y}}) = MSE(DCT(\mathbf{y}), DCT(\hat{\mathbf{y}})) \quad (3)$$

IV. Experiment and Results

A. Dataset and Training Details

Forty 3D head CT volumes from publicly available Mayo Clinic dataset [40] were used for the task. As the provided dataset has helical acquired projections, CTL tool [41] was used to create 496 cone-beam projections similar to a clinical setup. FDK was used for reconstruction, where each reconstructed head was of dimensionality $512 \times 512 \times 512$. For the low quality model input (LQ), only 15 low-dose projections were used for reconstruction as done in [42]. The high quality (HQ) groundtruth images were reconstructed using all (496) routine-dose projections.

The models were trained on the 2D slices of the CT volume. Therefore for each patient, there were 512 2D slices. Each 2D slice was cropped to 384×384, to retain only useful information. The image intensities were cropped to the 99th percentile and subsequently normalized to an interval of [0,1]. The dataset split for all the experiments was: training (32), validation (4) and test (4). We trained 4 models, using the same architecture (refer to Appendix B). Three models were trained on normal images. The fourth model (M_{input_dct}) used 8×8 block-wise DCT transformed input as in [31] and Loss_{mse} for all loss components. The model M_{mse} used Loss_{mse} similar to [9], which served as the baseline. The perception-aware loss models: M_{dct} used Loss_{dct}, M_{LoG} used Loss_{LoG}, for all the loss components. The other training details are mentioned in Appendix A.

B. Results and Discussion

1) Quantitative Results: We evaluate the performance of the models on test data using the following metrics, MSE, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM).

Model	SSIM	PSNR (dB)	MSE
FDK	0.176 $\pm$0.040	14.53 $\pm$1.94	3.92e-2 $\pm$0.020
M_{mse}	0.743 $\pm$0.072	24.35 $\pm$2.54	4.34e-3 $\pm$0.002
M_{dct}	**0.803** $\pm$0.071	**27.84**$\pm$2.41	**1.90e-3** $\pm$0.002
M_{LoG}	0.799 $\pm$0.066	27.11 $\pm$2.68	2.23e-3 $\pm$0.002
$\text{M}_{\text{input\_dct}}$	0.752 $\pm$0.067	26.00 $\pm$2.51	2.69e-3 $\pm$0.002

TABLE I
QUANTITATIVE EVALUATION OF THE RECONSTRUCTION QUALITY OF ALL MODELS AND FDK. VALUES ARE THE MEAN AND STANDARD DEVIATION OVER 2D SLICES. BEST RESULTS ARE MARKED BOLD.

Table I portrays that M_{dct} performed the best and M_{mse} the worst, with respect to all metrics. Comparing the three models trained on pixel space, we observe that the models using surrogate losses produced lower pixelwise MSE than the one trained with MSE objective. We can infer from the empirical results that perception-aware losses are beneficial for training, as they model pixel gradients instead of pixel values. The M_{input_dct} model also yields better performance than M_{mse}. This confirms

the hypothesis that DCT captures the pertinent details of the image which facilitates learning, as reported by other works [30, 31]. However, interestingly the model trained on image space and using loss $Loss_{dct}$ produced better results compared to M_{input_dct} trained in DCT space. The reason can be, the convolutional architecture is not suitable for this objective, as in the DCT domain identical patterns in different locations have different meaning.

2) Qualitative Results and Discussion: Fig. 3 shows that the models using perception-aware loss removed the streaking artifacts and reconstructed the high intensity details such as, bone and teeth fillings. For easier LQ images, such as Fig. 3 (d)-LQ, M_{mse} reconstructed the high intensity details. However, for difficult cases, such as Fig. 3 (c)-LQ, M_{mse} did not even remove the streaking artifacts completely. For some cases, the perception-aware loss based models did not reconstruct the soft-tissue well, as seen in Fig. 3 (c). We leave it as future work, where we consider to improve the soft-tissue contrast by incorporating the previous CT examinations as prior knowledge. For further analysis, we studied the progression of filter responses or activation maps over first few epochs for M_{mse}, M_{dct} and M_{LoG} models, where maximal change is expected. Generally for lower level filters, the pixels were connected and formed structure much earlier for M_{dct} and M_{LoG} models compared to M_{mse}. Similarly, for high-level filters, we observed that the inner structural details were prominent at a much earlier epoch for M_{dct} and M_{LoG} models, compared to M_{mse} (refer to Fig 4). Further, we can see the edges seem more highlighted for LoG, attributed to the Laplacian operator's edge enhancement properties.

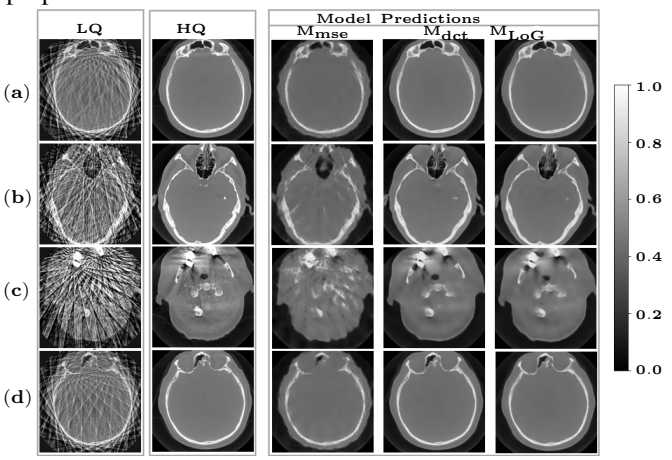

Fig. 3. The predictions produced by the models for 4 cases from test data. 1st column contains LQ (Input), 2nd column contains the corresponding HQ (Groundtruth), and the rightmost pane contains predictions from models M_{mse}, M_{dct} and M_{LoG} respectively for the corresponding LQs.

3) Analysis of the Role of Embedding Levels: Our architecture has three embedding levels as shown in Fig. 2. We tracked the progression of activation maps for both LQ and its corresponding HQ image. We inspect the epochwise activation maps for both LQ and HQ, to investigate what each embedding level captures. Fig. 5 (a) portrays that for top-most embedding Q1, the activation maps of LQ are evolving to develop star-like patterns (artifacts), whereas the HQ activation maps evolved without artifacts. This indicates that the Q1 still encodes the artifacts, and the artifact removal is also done by the decoder. Fig. 5 (b) shows that the mid-level embedding Q2 captured more global structures than Q1, such as the geometry of the head and inner bone or tissue details. Further, Q2 captures more local details than the lowest level embedding Q3, as seen in Fig. 5 (c). Further, we can observe that the HQ and LQ activation maps tend to become similar over the epochs, for both Q2 and Q3. The empirical results support that each embedding level of VQ-VAE captures distinguishable concepts, as we intended with our architecture design. This observation was consistent for different examples.

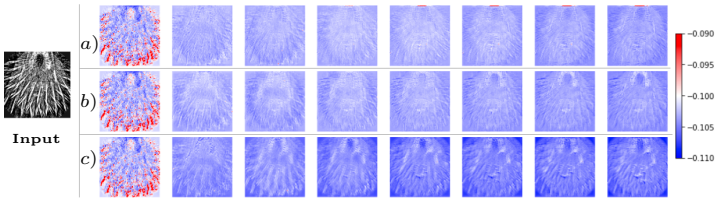

Fig. 4. The progression of the activation maps for the input image (left side) for first 7 epochs. The activation maps are produced by models: (a) M_{mse} (b) M_{dct} (c) M_{LoG}.

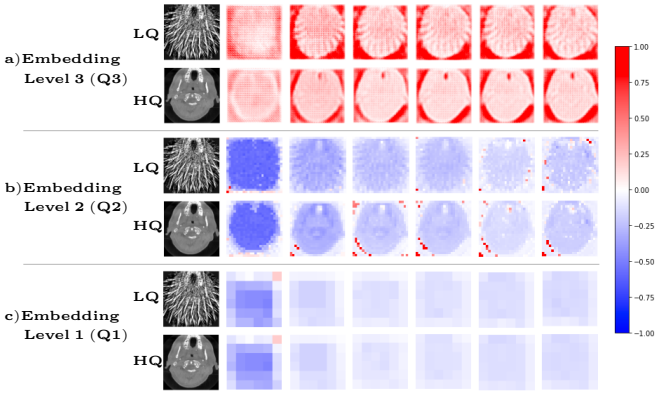

Fig. 5. The progression of the activation maps for LQ and HQ for first 6 epochs. The activation maps are produced from the best performing model M_{dct}, from the respective embedding levels (a) top-most Q1 (b) middle Q2 (c) lowest Q3

V. Conclusion

In this work, we introduced two simple but effective perception-aware MSE based losses. The losses enable the hierarchical VQ-VAE framework to remove the challenging low-dose related artifacts. Further, they help the model to capture the fine details (bones, teeth and metal fillings) accurately compared to the pixelwise MSE model. The losses are not architecture specific and easy to implement. We also provided an analysis on the role of different embedding levels. The analysis supported that each level captured distinct concepts at different abstraction levels. This facilitated the removal of the artifacts in the low quality head CT images, which were generated from only 3% (15 out of 496) of the projections.

Appendix A
Experiment and Hyperparameter Details

The models are implemented in Tensorflow(2.3.0). The experiments were carried out on NVIDIA Tesla V100 32GB GPUs. Each experiment was run for maximum of 500 epochs with early stopping, where training was stopped if there was no improvement after 20 epochs. Online data augmentation included random flipping, rotation and scaling. The models were trained using ADAM [43] optimiser using learning rate

1e-4. The model with the lowest validation SSIM was selected and used for the evaluation on test data. The batch size was set to 72 (the maximum which fits the GPU memory) for all experiments. We used type-II variant of DCT,same as used in [29]. The Laplacian operator (refer to Equation 4) used in the LoG loss is implemented as a discrete convolution kernel, [1,1,1; 1,-8,1; 1,1,1]. The parameters for Gaussian kernel were set as sigma=1.0 and kernel size=7. The LeakyReLU slope was set as 0.2.

$$\Delta^2 f(x,y) = \underbrace{f(x+1,y)+f(x-1,y)-2f(x,y)}_{x\ direction} + \underbrace{f(x,y+1)+f(x,y-1)-2f(x,y)}_{y\ direction} + \underbrace{f(x-1,y-1)+f(x-1,y+1)+f(x+1,y+1)+f(x+1,y-1)+4f(x,y)}_{diagonal} \tag{4}$$

APPENDIX B
VQ-VAE ARCHITECTURE DIAGRAM

Fig. 6 shows the detailed hierarchical VQ-VAE architecture implemented for the task. The three embedding levels are marked as Q1, Q2 and Q3. The Residual+strided convolution block comprises of two residual blocks followed by a strided convolution having kernel size 3×3 and stride of two. The residual block is implemented as: activation, 3×3 convolution, activation, 3×3 convolution, where LeakyReLU [44] was chosen for the activation. The decoder similarly has two residual blocks, followed by one transposed convolution with stride of two and kernel size 3×3.

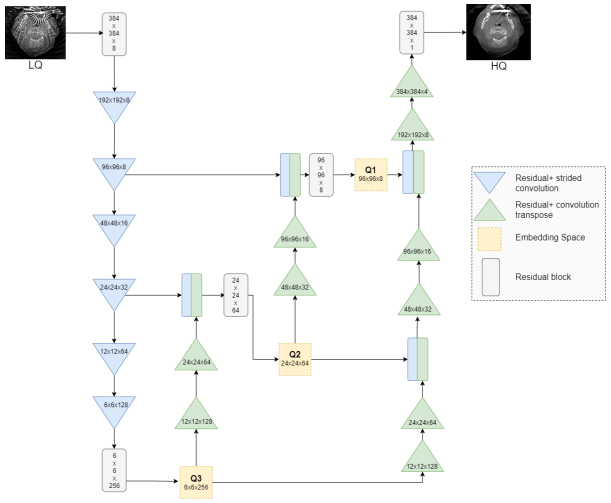

Fig. 6. The hierarchical VQ-VAE architecture for the CT denoising and artifact removal task.

ACKNOWLEDGMENT

This work was conducted within the context of the International Graduate School MEMoRIAL at OVGU Magdeburg, Germany, kindly supported by the European Structural and Investment Funds (ESF) under the programme "Sachsen-Anhalt WISSENSCHAFT Internationalisierung" (project number ZS/2016/08/80646). The research was further supported through the project "CogXAI – Cognitive neuroscience inspired techniques for eXplainable AI" funded by the Federal Ministry of Education and Research of Germany (BMBF). The funding agencies had no role in the study design, decision to publish, or preparation of this manuscript.

REFERENCES

[1] *OECD/European Union (2020), Health at a Glance: Europe 2020: State of Health in the EU Cycle*. OECD Publishing, Paris, 2020.

[2] R. Gupta, C. Walsh, I. S. Wang, M. Kachelrieß, J. Kuntz, and S. Bartling, "Ct-guided interventions: current practice and future directions," *Intraoperative Imaging and Image-Guided Therapy*, pp. 173–191, 2014.

[3] C. G. Macias and J. J. Sahouria, "The appropriate use of ct: quality improvement and clinical decision-making in pediatric emergency medicine," *Pediatric radiology*, vol. 41, no. 2, pp. 498–504, 2011.

[4] M. J. Willemink and P. B. Noël, "The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence," *European radiology*, vol. 29, no. 5, pp. 2185–2195, 2019.

[5] K. Mei, F. K. Kopp, R. Bippus, T. Köhler, B. J. Schwaiger, A. S. Gersing, A. Fehringer, A. Sauter, D. Münzel, F. Pfeiffer *et al.*, "Is multidetector ct-based bone mineral density and quantitative bone microstructure assessment at the spine still feasible using ultra-low tube current and sparse sampling?" *European radiology*, vol. 27, no. 12, pp. 5261–5271, 2017.

[6] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *Josa a*, vol. 1, no. 6, pp. 612–619, 1984.

[7] H.-M. Zhang and B. Dong, "A review on deep learning in medical image reconstruction," *Journal of the Operations Research Society of China*, pp. 1–30, 2020.

[8] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[9] A. Razavi, A. v. d. Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *arXiv preprint arXiv:1906.00446*, 2019.

[10] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep admm-net for compressive sensing mri," in *Proceedings of the 30th international conference on neural information processing systems*, 2016, pp. 10–18.

[11] X. Zheng, S. Ravishankar, Y. Long, and J. A. Fessler, "Pwls-ultra: An efficient clustering and learning-based approach for low-dose 3d ct image reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1498–1510, 2018.

[12] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "Cnn-based projected gradient descent for consistent ct image reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1440–1453, 2018.

[13] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[14] S. Bazrafkan, V. Van Nieuwenhove, J. Soons, J. De Beenhouwer, and J. Sijbers, "Deep learning based computed tomography whys and wherefores," *arXiv preprint arXiv:1904.03908*, 2019.

[15] L. Shen, W. Zhao, and L. Xing, "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning," *Nature*

biomedical engineering, vol. 3, no. 11, pp. 880–888, 2019.

[16] Y. Kasten, D. Doktofsky, and I. Kovler, "End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images," in *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer, 2020, pp. 123–133.

[17] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, "A sparse-view ct reconstruction method based on combination of densenet and deconvolution," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1407–1417, 2018.

[18] S. Xie, X. Zheng, Y. Chen, L. Xie, J. Liu, Y. Zhang, J. Yan, H. Zhu, and Y. Hu, "Artifact removal using improved googlenet for sparse-view ct reconstruction," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[19] T. Humphries, D. Si, S. Coulter, M. Simms, and R. Xing, "Comparison of deep learning approaches to low dose ct using low intensity and sparse view data," in *Medical Imaging 2019: Physics of Medical Imaging*, vol. 10948. International Society for Optics and Photonics, 2019, p. 109484A.

[20] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose ct via convolutional neural network," *Biomedical optics express*, vol. 8, no. 2, pp. 679–694, 2017.

[21] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.

[22] D. Lee, S. Choi, and H.-J. Kim, "High quality imaging from sparsely sampled computed tomography data with deep learning and wavelet transform in various domains," *Medical physics*, vol. 46, no. 1, pp. 104–115, 2019.

[23] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1740–1749.

[26] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 679–691, 2005.

[27] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting gan dct anomalies," *arXiv preprint arXiv:2101.09781*, 2021.

[28] Y. Wang, C. Xu, C. Xu, and D. Tao, "Packing convolutional neural networks in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 10, pp. 2495–2510, 2018.

[29] M. Ulicny, V. A. Krylov, and R. Dahyot, "Harmonic convolutional networks based on discrete cosine transform," *arXiv preprint arXiv:2001.06570*, 2020.

[30] M. Ulicny and R. Dahyot, "On using cnn with dct based image data," in *Proceedings of the 19th Irish Machine Vision and Image Processing conference IMVIP*, vol. 2, 2017.

[31] L. Gueguen, A. Sergeev, B. Kadlec, R. Liu, and J. Yosinski, "Faster neural networks straight from jpeg," *Advances in Neural Information Processing Systems*, vol. 31, pp. 3933–3944, 2018.

[32] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics*, vol. 4, 2007.

[33] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3d-dct based perceptual quality assessment of stereo video," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 2521–2524.

[34] D. J. Rezende and F. Viola, "Taming vaes," *arXiv preprint arXiv:1810.00597*, 2018.

[35] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," 2019.

[36] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[37] W. Xue, X. Mou, L. Zhang, and X. Feng, "Perceptual fidelity aware mean squared error," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 705–712.

[38] C. Maccone, "A simple introduction to the klt (karhunen—loève transform)," *Deep space flight and communications: exploiting the sun as a gravitational lens*, pp. 151–179, 2009.

[39] R. C. Gonzalez, R. E. Woods *et al.*, "Digital image processing," 2002.

[40] C. McCollough, B. Chen, D. Holmes, X. Duan, Z. Yu, L. Xu, S. Leng, and J. Fletcher, "Low dose ct image and projection data [data set]," *The Cancer Imaging Archive*, 2020.

[41] T. Pfeiffer, R. Frysch, R. N. Bismark, and G. Rose, "Ctl: modular open-source c++-library for ct-simulations," in *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, vol. 11072. International Society for Optics and Photonics, 2019, p. 110721L.

[42] F. Saad, R. Frysch, V. Kulvait, D. Punzet, and G. Rose, "Nullspace-constrained modifications of under-sampled interventional ct images using instrument-specific prior information," in *Medical Imaging 2020: Physics of Medical Imaging*, vol. 11312. International Society for Optics and Photonics, 2020, p. 113123J.

[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015.

[44] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30. Citeseer, 2013, p. 3.