2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Cascaded Learning with Generative Adversarial Networks for Low Dose CT Denoising

Sepehr Ataei[1], Paul Babyn[2], Alireza Ahmadian[3], Javad Alirezaie[1*]

***Abstract*—CT machines can be tuned in order to reduce the radiation dose used for imaging, yet reducing the radiation dose results in noisy images which are not suitable in clinical practice. In order for low dose CT to be used effectively in practice this issue must be addressed. Generative Adversarial Networks (GAN) have been used widely in computer vision research and have proven themselves as a powerful tool for producing images with high perceptual quality. In this work we use a cascade of two neural networks, the first is a Generative Adversarial Network and the second is a Deep Convolutional Neural Network. The first network generates a denoised sample which is then fine-tuned by the second network via residue learning. We show that our cascaded method outperforms related works and more effectively reconstructs fine structural details in low contrast regions of the image.**

***Index Terms*—Image Reconstruction; Computed Tomography; Computer Vision; Convolutional Neural Network**

I. Introduction

Radiation exposure is a significant risk associated with X-ray computed tomography (CT). Excessive radiation exposure can have negative side effects and may cause cancer. However reducing radiation dose also reduces the signal to noise ratio (SNR) which may impact diagnostic accuracy. The research community has thus proposed various noise reduction techniques in order to denoise LDCT images such that they may be used instead of Normal Dose CT (NDCT). This has been challenging because the noises present in LDCT images are non-Gaussian and spatially variant and thus not well defined. Deep Learning and Deep Convolutional Neural Networks (DCNNs) have recently emerged as a powerful computer vision technique capable of approximating unknown statistical distributions based on training samples. As such, DCNNs appear to be an appropriate tool for solving the LDCT denoising problem. DCNNs have achieved state-of-the-art in many computer vision and image processing tasks [1]. In medical imaging, these networks have outperformed other methods in segmentation [2], classification [3], denoising [4] and more. Early work focused on development of iterative reconstructive techniques (IR) algorithms for Low Dose CT (LDCT) image reconstruction, however; these algorithms are computationally expensive and may cause artifacts in CT images. These methods require access to raw projection data and as a result are not compatible across devices from different manufacturers. Image post-processing techniques such as DCNNs on the other hand do not require projection data and thus offer a more general and scanner independent solution.

In the past few years, the design of efficient DCNN architectures for LDCT denoising has been researched and several approaches proposed. Ansari et. al. proposed a Dilated Residual Learning (DRL) model architecture [5] which achieved good performance with a relatively small network.This design used two distinct features. First, they used the benefits of dilated convolutions for increasing the receptive field of the network without increasing the parameter count and depth. Second, they used residual connections [1] between layers to allow for information from shallow layers to propagate to deeper layers. Various GAN (Generative Adversarial Network) based approaches have also been proposed [6], [7]. Yang et al. [6] used Wasserstein GAN and successfully generated normal dose CT (NDCT) from LDCT. GANs are able to learn a target distribution and then generate new samples which look visually similar to samples in the target distribution [7]. This makes GANs a powerful tool for modelling the human visual system. Wu et. al. proposed that instead of increasing the complexity and depth of networks, one can use a cascade of simple CNNs [8]. Significant research has also been done on the design of loss functions for optimizing DCNNs. It quickly became evident that mean-squared-error (MSE) tends to over-smooth images and cause the loss of fine structural details. Researchers proposed linear combinations of MSE and other loss terms such as Perceptual Loss [6] and Structural Dissimilarity [9] in order to overcome this limitation. Calculating the perceptual loss involves using a pre-trained network as a feature extractor. These pre-trained networks (such as VGG19 [10]) have been trained on very large datasets for image classification (ImageNet) and thus can be used to model the human visual system.

While perceptual loss has proven to be an effective loss term in order to promote the perceptual similarity of the images [6], we show that it can introduce patterned artifacts into the image. These artifacts can lead to miss-diagnosis by radiologists and should be avoided. Further, computing the perceptual loss introduces significant additional computational load including increased training time and memory requirements.

In this work, we propose a new cascaded network structure (Figure 2). We use MSE to train a Least Squares GAN (LSGAN) at the first level as a rough estimate of the target. We show that our GAN approach is able to maintain texture and the perceptual quality of images. In the next level, we train the DRL network to fine tune the result by predicting the

[1]Department of Electrical and Computer Engineering, Ryerson Univeristy, Toronto, ON M5B2K3 Canada (s2ataei@ryerson.ca; javad@ryerson.ca); [2] Department of Medical Imaging, University of Saskatoon, Saskatoon, SK S7N0W8 Canada (Paul.Babyn@saskhealthauthority.ca); [3] Tehran University of Medical Sciences, Tehran, Iran (ahmadian@sina.tums.ac.ir), *corresponding author: J Alirezaie

difference between ground truth normal dose CT images and denoised predictions from the GAN. We show that the results of this cascade of networks outperforms each component network. Lastly, we show that our approach is more effective than perceptual based approaches and produces images with high peak signal to noise ratio (PSNR) and structural similarity (SSIM) scores in addition to high perceptual quality.

II. Background

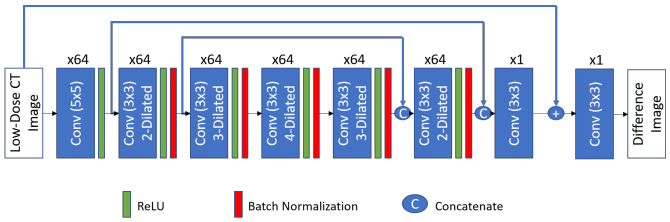

Fig. 1. DRL Model Architecture

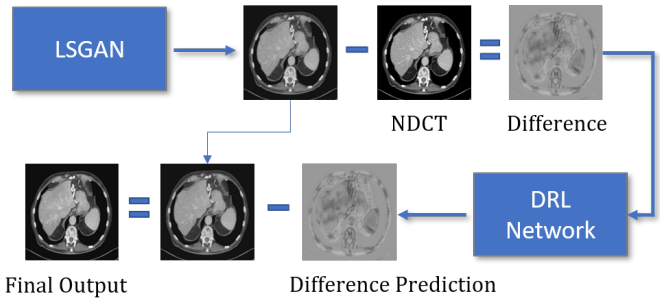

Fig. 2. Proposed Cascaded System Block Diagram

A. Residual Learning

Adding additional layers to neural networks has been shown to improve performance by aiding the optimization of a larger feature space, however; research has shown that there is an upper bound on this benefit. As the number of layers reaches this upper bound, training becomes problematic due to the vanishing and/or exploding gradient problem. To overcome this, He et. al. proposed residual networks. They show that simply increasing the number of layers in a network is not an effective method to improve model accuracy as accuracy will degrade past a certain depth [1]. As a solution they proposed a residual learning process during which the input to a residual block is added to the output of the same block which may contain any number of traditional layers. The output of the residual block is then used as input to the next residual block in a chain. This allows information from early layers in the network to more easily propagate to the deeper layers and allows the network to learn low level, medium level and high level features.

B. Generative Adversarial Networks

GANs were introduced by Goodfellow et al. in 2014 [11]. The architecture includes a generator network which learns the desired data distribution G and a discriminator network which estimates the probability that a sample came from the training data rather than G. While training, the networks learn simultaneously as the discriminator improves at distinguishing real samples from fake samples, the generator must then learn to fool the discriminator. This causes the generator to produce better and better samples.

Modified GAN approaches have been developed in order to stabilize the training and avoid gradient based problems. In this work, we use the Least Squares Generative Adversarial Network (LSGAN) proposed by Mao et. al. [12].

$$\begin{aligned}\min_D V_{\text{LSGAN}}(D) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x} \mid \Phi(\boldsymbol{y})) - 1)^2\right] \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}) \mid \Phi(\boldsymbol{y})))^2\right] \\ \min_G V_{\text{LSGAN}}(G) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}) \mid \Phi(\boldsymbol{y})) - 1)^2\right].\end{aligned} \tag{1}$$

C. Perceptual Loss

Yang et al. define a perceptual loss based on features extracted by the pre-trained VGG-19 network [6]. They use the output of the 16th convolutional layer as the extracted features, and calculate a mean squared error distance between the features of the ground truth image and corresponding denoised image. The perceptual loss $L_p(\theta)$ is defined in Equation 2. $\hat{y}(\theta)$ is a denoised image and y is the corresponding ground truth image. Feature maps ϕ_i are extracted from block i of the pre-trained VGG19 network with size $h_i \times w_i \times d_i$

$$L_P(\theta) = \sum_{i=1}^{4} \frac{1}{h_i w_i d_i} \left\|\phi_i(\hat{y}(\theta)) - \phi_i(y)\right\|^2. \tag{2}$$

Perceptual loss can be effective for denoising because it more accurately models the human visual system than MSE. MSE only compares per pixel difference between two images and does not take into account high level features. Deep CNNs such as VGG-19 are better able to model the human visual system because they learn features to accurately describe the natural images they are trained on [13]. One can take advantage of this artificial visual understanding by penalizing a network when extracted features are dissimilar. Based on this justification, many works have used perceptual loss for low dose CT denoising.

In this work, we found that perceptual loss can have a negative impact on low dose CT denoising as it can lead to patterned artifacts. Further, using deep networks such as VGG19 as a feature extractor introduces significant memory requirements and additional computation time during training. Therefore, we propose the utilization of GANs for producing images with high perceptual quality in the first level, and fine tune these images in the second level to the final denoised images.

III. Methods

A. Denoising Model

We can represent an image denoiser G mathematically as a function that maps LDCT to NDCT ,where $\boldsymbol{z} \in \mathbb{R}^{N \times N}$ is a LDCT image and $\boldsymbol{x} \in \mathbb{R}^{N \times N}$ is a NDCT image.

$$G : z \rightarrow x \tag{3}$$

Although noise in raw x-ray measurements can be modelled as a combination of Poisson quantum noise and Gaussian electronic noise, reconstructed CT images do not have a well-defined noise distribution and the noise is non-uniformly distributed across the image. Deep neural networks are advantageous here as they can efficiently learn high-level features and accurate data representations given a sufficiently large training set.

B. Dataset Preparation

We train and evaluate our methods on the AAPM Low Dose CT Grand Challenge dataset. We extracted 281,660 patches of size 64x64 from 8 patients for training, and used the remaining 844 full sized (512 x 512) images from the remaining two patients for testing. Due to the fully convolutional nature of the proposed networks we are able to train on patches and test on full sized images. Patches help to reduce memory requirements and increase the number of training samples.

C. Objective Functions

For MSE we minimize the loss function L with respect to model parameters θ where $\left\{(x_i, y_i)\right\}_{i=1}^{N}$ are NDCT and LDCT image pairs respectively.

$$L_{MSE}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| f\left(x_i; \theta\right) - y_i \right\|_F^2 \tag{4}$$

The DRL network is trained using Equation 4 and the LSGAN network is trained using Equation 1.

D. Cascade Architecture

Figure 2 gives a breakdown of the proposed architecture. In the first level of the cascade, the LSGAN network inputs are LDCT and the labels are NDCT images where $\mathbf{X}$ is the LDCT image and $\mathbf{Y}$ is the NDCT image. In the second level of the cascade, the DRL network (Figure 1) inputs are predictions from the first level $\mathbf{Y}_P^{(1)}$ and the labels are the difference image between first level predictions and NDCT images $\mathbf{Y}_P^{(2)} = \mathbf{Y}_P^{(1)} - \mathbf{Y}$.

During testing, LDCT images are given as input to the LSGAN which gives preliminary denoised images as output. These images are then input to the DRL network which outputs predicted difference images $\mathbf{Y}_P^{(2)}$. To arrive at the final output, we subtract $\mathbf{Y}_P^{(1)}$ from $\mathbf{Y}_P^{(2)}$ to arrive at the final prediction $\mathbf{Y}_{final}$.

IV. Experiments and Results

For classification and segmentation there exists accurate quantitative scores used to compare methods on the same data. For denoising however, the most commonly used quantitative metrics are PSNR and SSIM. These metrics while powerful, have been shown to be insufficient. Minimizing MSE will maximize PSNR mathematically, but research has shown that simply maximizing MSE can over-smooth images [6], [5]. Evaluating LDCT denoising methods cannot rely solely on current quantitative metrics, and researchers have turned to qualitative studies with radiologists to score the perceptual quality of images. Groups that do not have access to radiologists for evaluation must perform some qualitative analysis on their own.

The LDCT Grand Challenge Dataset contains annotations for CT slices containing lesions. These slices are important because in practice radiologists are looking for lesions during CT examinations. In contrast to other works which select the most desirable testing examples for performing qualitative analysis, we avoid cherry-picking and perform qualitative analysis on testing samples which have been pre-annotated with lesions. This method of results evaluation is less biased and gives a more honest representation of model performance.

TABLE I
Mean PSNR and SSIM for test set. The best and the second best are denoted in red and blue, respectively.

Method	PSNR	SSIM
Low Dose	25.4247	0.8211
BM3D	26.8496	0.7122
FC-AIDE [14]	29.1620	0.8804
DRLP [5]	28.8069	0.8436
CCNN-PL [4]	29.7232	0.8799
Proposed	**29.2404**	**0.8816**

V. Discussion and Conclusion

Table I shows the mean PSNR and SSIM scores for our proposed method compared to other popular denoising models. Our method achieves the best SSIM score and the second best PSNR score. As previously mentioned, minimizing MSE will lead to maximized PSNR however images may be over-smoothed and lose fine structural and textural details. Achieving both high PSNR and SSIM scores however indicates a well performing model which has removed the noise and maintained structural details. Our proposed method demonstrates this quality quantitatively and this finding is further supported by our qualitative analysis.

Figure 3 demonstrates the proposed cascaded structures ability to fine tune results and achieve better performance than each of the cascade layers independently. To re-iterate, LSGAN is the output from the first cascade level and DRL is the second cascade level. The proposed image shows the using those two levels in series and demonstrates that there is visible improvements by cascading the two.

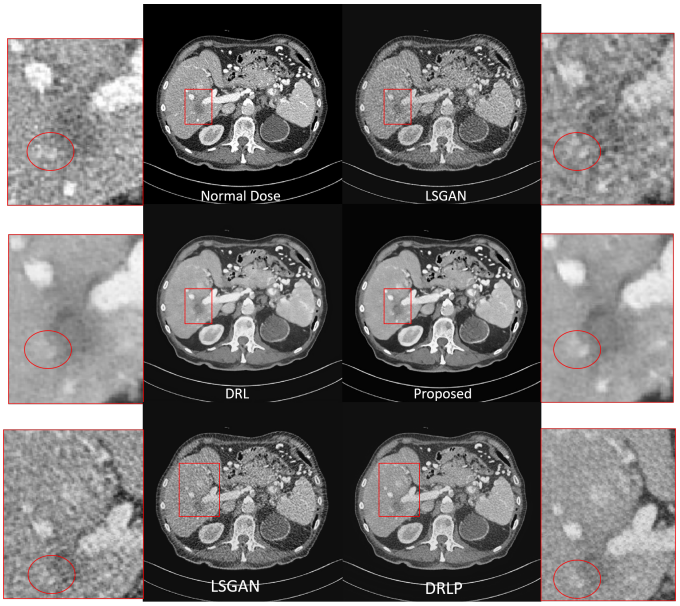

Fig. 3. Proposed Method Denoising Results

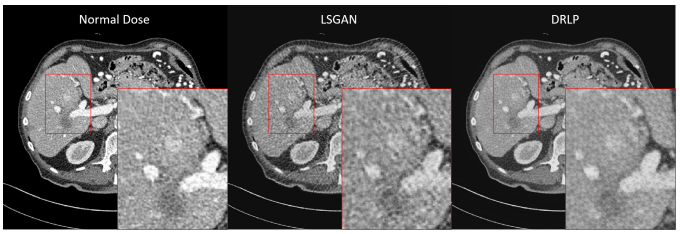

Fig. 4. Comparing texture and perceptual quality of GAN and perceptual loss

Figure 4 highlights the shortcomings of perceptual loss at modeling accurate structural and textural details during denoising. It is evident from the DRLP results (DRL method trained with perceptual loss) that perceptual loss is introducing a patterned artifact into the denoising result. The network has demonstrated some ability and understanding of the background texture however it has not accurately reconstructed this texture. This background texture inherent to CT images results from spatially variant noises. By definition, a patterned artifact will not accurately model these spatially variant noises effectively. Comparing the DRLP prediction with the LSGAN prediction demonstrates that the LSGAN is able to more effectively model the spatially variant noises as the texture is not patterned. This finding suggests that GANs may be a more effective tool when trying to maintain perceptual quality of CT images during denoising compared to perceptual loss. Further, calculating perceptual loss increases training time and memory requirements significantly.

In this paper we proposed a cascade of deep neural networks in order to denoise LDCT images. The first network was LSGAN, and the second network was DRL. Input images were initially denoised by LSGAN and subsequently fine tuned by the DRL network. The DRL network predicted the difference image, which was subtracted from the LSGAN predictions to produce the final denoised output. Our proposed method achieved high PSNR and perceptual quality without introducing a perceptual loss. It was explained that perceptual loss greatly enhances the ability of the model to preserve structural details during denoising. However, perceptual loss introduces a significant increase in memory requirements and training time. Our poposed method was able to maintain perceptual quality of the images by using LSGAN at the first cascade level instead of relying on a perceptual loss. We introduced residue image based denoising in the second level to fine tune our results and improve the PSNR and SSIM scores.

In conclusion, we have proposed an architecture which outperforms the related works while being less computationally expensive in testing and training. We have proven that our cascaded architecture is able to produce denoised images with better perceptual quality and relatively high PSNR and SSIM scores when compared to other MSE based methods.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[2] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[3] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[4] Sepehr Ataei, J. Alirezaie, and P. Babyn. Cascaded convolutional neural networks with perceptual loss for low dose ct denoising. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5, 2020.

[5] Maryam Gholizadeh-Ansari, Javad Alirezaie, and Paul S. Babyn. Deep learning for low-dose ct denoising using perceptual loss and edge detection layer. *Journal of Digital Imaging*, pages 1 – 12, 2019.

[6] 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*, 37(6):1348–1357, June 2018.

[7] Xin Yi and Paul Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of Digital Imaging*, 31, 08 2017.

[8] Dufan Wu, Kyung Sang Kim, Georges El Fakhri, and Quanzheng Li. A cascaded convolutional nerual network for x-ray low-dose ct image denoising. *ArXiv*, abs/1705.04267, 2017.

[9] S. Ataei, J. Alirezaie, and P. Babyn. Low dose ct denoising using dilated residual learning with perceptual loss and structural dissimilarity. In *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME)*, pages 1–5, 2020.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[12] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017.

[13] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. volume 9906, pages 694–711, 10 2016.

[14] S. Cha and T. Moon. Fully convolutional pixel adaptive image denoiser. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4159–4168, 2019.