Noise Conditioned Weight Modulation for Robust and Generalizable Low Dose CT Denoising

Sutanu Bera(✉) and Prabir Kumar Biswas

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
sutanu.bera@iitkgp.ac.in

Abstract. Deep neural networks have been extensively studied for denoising low-dose computed tomography (LDCT) images, but some challenges related to robustness and generalization still need to be addressed. It is known that CNN-based denoising methods perform optimally when all the training and testing images have the same noise variance, but this assumption does not hold in the case of LDCT denoising. As the variance of the CT noise varies depending on the tissue density of the scanned organ, CNNs fails to perform at their full capacity. To overcome this limitation, we propose a novel noise-conditioned feature modulation layer that scales the weight matrix values of a particular convolutional layer based on the noise level present in the input signal. This technique creates a neural network that is conditioned on the input image and can adapt to varying noise levels. Our experiments on two public benchmark datasets show that the proposed dynamic convolutional layer significantly improves the denoising performance of the baseline network, as well as its robustness and generalization to previously unseen noise levels.

Keywords: LDCT denoising · Dynamic Convolution · CT noise variance

1 Introduction

Convolutional neural networks (CNN) have emerged as one of the most popular methods for noise removal and restoration of LDCT images [1,2,5,6,14]. While CNNs can produce better image quality than manually designed functions, there are still some challenges that hinder their widespread adoption in clinical settings. Convolutional denoisers are known to perform best when the training and testing images have similar or identical noise variance [15,16]. On the other hand, different anatomical sites of the human body have different tissue densities and compositions, which affects the amount of radiation that is absorbed and scattered during CT scanning; as a result, noise variance in LDCT images also varies significantly among different sites of the human body [13].

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 88–97, 2023.
https://doi.org/10.1007/978-3-031-43999-5_9

Furthermore, the noise variance is also influenced by the differences in patient size and shape, imaging protocol, etc. [11]. Because of this, CNN-based denoising networks fail to perform optimally in LDCT denoising. In this study, we have introduced a novel dynamic convolution layer to combat the issue of noise level variability in LDCT images. Dynamic convolution layer is a type of convolutional layer in which the convolutional kernel is generated dynamically at each layer based on the input data [3,4,8]. Unlike the conventional dynamic convolution layer, here we have proposed to use a modulating signal to scale the value of the weight vector(learned via conventional backpropagation) of a convolutional layer. The modulating signal is generated dynamically from the input image using an encoder network. The proposed method is very simple, and learning the network weight is a straightforward one-step process, making it manageable to deploy and train. We evaluated the proposed method on the recently released large-scale LDCT database of TCIA Low Dose CT Image and Projection Data [10] and the 2016 NIH-AAPM-Mayo Clinic low dose CT grand challenge database [9]. These databases contain low-dose CT data from three anatomical sites, i.e., head, chest, and abdomen. Extensive experiments on these databases validate the proposed method improves the baseline network's performance significantly. Furthermore, we have shown the generalization ability to the out-of-distribution data, and the robustness of the baseline network is also increased significantly via using the proposed weight-modulated dynamic convolutional layer.

2 Method

Motivation: Each convolutional layer in a neural network performs the sum of the product operation between the weight vector and input features. However, as tissue density changes in LDCT images, the noise intensity also changes, leading to a difference in the magnitude of intermediate feature values. If the variation in input noise intensity is significant, the magnitude of the output feature of the convolutional layer can also change substantially. This large variation in input feature values can make the CNN layer's response unstable, negatively impacting the denoising performance. To address this issue, we propose to modulate the weight vector values of the CNN layer based on the noise level of the input image. This approach ensures that the CNN layer's response remains consistent, even when the input noise variance changes drastically.

Weight Modulation: Figure 1 depicts our weight modulation technique, which involves the use of an additional anatomy encoder network, $\mathcal{E}_a$, along with the backbone denoising network, CNN_D. The output of the anatomy encoder, denoted as e_x, is a D-dimensional embedding, i.e., $e_x = \mathcal{E}_a(\nabla^2(x))$. Here, x is the input noisy image, and $\nabla^2(.)$ is a second-order Laplacian filter. This embedding e_x serves as a modulating signal for weight modulation in the main denoising network (CNN_D). Specifically, the *lth* weight-modulated convolutional layer, $\mathcal{F}_l$, of the backbone network, CNN_D, takes the embedding e_x as input. Then the embedding e_x is passed to a 2 Layer MLP, denoted as ϕ_l, which learns a non-linear mapping between the layer-specific code, denoted as $s_l \in \mathbb{R}^{N_l}$, and the

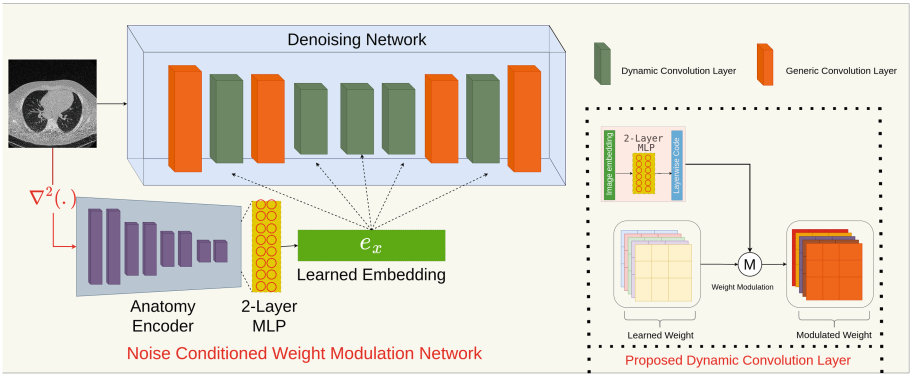


Fig. 1. Overview of the proposed noise conditioned weight modulation framework.

embedding e_x, i.e., $s_l = \phi_l(e_x)$. Here, N_l represents the number of feature maps in the layer $\mathcal{F}_l$. The embedding e_x can be considered as the high dimensional code containing the semantics information and noise characteristic of the input image. The non-linear mapping ϕ_l maps the embedding e_x to a layer-specific code s_l, so that different layers can be modulated differently depending on the depth and characteristic of the features. Let $w_l \in \mathbb{R}^{N_l \times N_{l-1} \times k \times k}$ be the weight vector of $\mathcal{F}_l$ learned via standard back-propagation learning. Here $(k \times k)$ is the size of the kernel, N_{l-1} is the number of feature map in the previous layer. Then the w_l is modulated using s_l as following,

$$\hat{w}_l = w_l \odot s_l \tag{1}$$

Here, $\hat{w}_l$ is the modulated weight value, and $\odot$ represents component wise multiplication. Next, the scaled weight vector is normalized by its L2 norm across channels as follows:

$$\tilde{w}_l = \hat{w}_l \Bigg/ \sqrt{\sum_{N_{l-1},k,k} \hat{w}_l^2 + \epsilon} \tag{2}$$

Normalizing the modulated weights takes care of any possible instability arise due to high or too low weight value and also ensures that the modulated weight has consistent scaling across channels, which is important for preserving the spatial coherence of the denoised image [7]. The normalized weight vectors, $\tilde{w}_l$ are then used for convolution, i.e., $f_l = \mathcal{F}_l\big(\tilde{w}_l * f_{l-1}\big)$. Here, f_l, and f_{l-1} are the output feature map of lth, $l-1$th layer, and $*$ is the convolution operation.

Relationship with Recent Methods: The proposed weight modulation technique leveraged the recent concept of style-based image synthesis proposed in StyleGAN2 [7]. However, StyleGAN2 controlled the structure and style of the generated image by modulating weight vectors using random noise and latent code. Whereas, we have used weight modulation for dynamic filter generation conditioned on input noisy image to generate a consistent output image.

Implementation Details: The proposed dynamic convolutional layer is very generic and can be integrated into various backbone networks. For our denoising task, we opted for the encoder-decoder-based UNet [12] architecture and replaced some of its generic convolutional layers with our weight-modulated dynamic convolution layer. To construct the anatomy encoder network, we employed ten convolutional blocks and downscaled the input feature map's spatial resolution by a factor of nine through two max-pooling operations inside the network. We fed the output of the last convolutional layer into a global average pooling layer to generate a 512-dimensional feature vector. This vector was then passed through a 2-layer MLP to produce the final embedding, $e_x \in \mathbb{R}^{512}$.

3 Experimental Setting

We used two publicly available data sets, namely, 1. TCIA Low Dose CT Image and Projection Data, 2. 2016 NIH-AAPM-Mayo Clinic low dose CT grand challenge database to validate the proposed method. The first dataset contains LDCT data of different patients of three anatomical sites, i.e., head, chest, and abdomen, and the second dataset contains LDCT images of the abdomen with two different slice thicknesses (3 mm, 1 mm). We choose 80% data from each anatomical site for training and the remaining 20% for testing. We used the Adam optimizer with a batch size of 16. The learning rate was initially set to $1e^{-4}$ and was assigned to decrease by a factor of 2 after every 6000 iterations.

Table 1. Objective and computational cost comparison between different methods. Objective metrics are reported by averaging the values for all the images present in the test set.

Model	Abdomen			Head			Chest			FLOPs
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE	
M1	33.84	0.912	8.46	39.45	0.957	2.42	29.39	0.622	103.27	75.53G
M2	34.15	0.921	7.41	40.04	0.968	2.02	29.66	.689	89.23	98.47G

4 Result and Discussion

Comparison with Baseline: This section discusses the efficacy of the proposed weight modulation technique, comparing it with a baseline UNet network (M1) and the proposed weight-modulated convolutional network (M2). The networks were trained using LDCT images from a single anatomical region and tested on images from the same region. Table 1 provides an objective comparison between the two methods in terms of PSNR, SSIM, and RMSE for different anatomical regions. The results show that the proposed dynamic weight modulation technique significantly improved the denoising performance of the baseline UNet for all settings. For example, the PSNR for head images was improved by 0.59 dB, and similar improvements were observed for other anatomical regions. Additionally, Table 1 shows the floating point computational requirements of the different

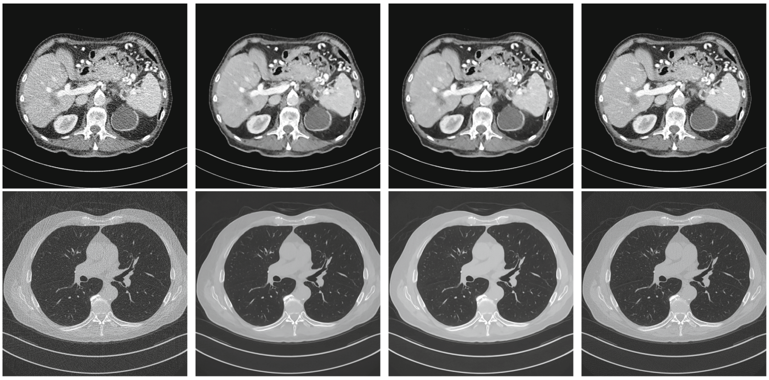

(a) LDCT (b) M1 (c) M2 (d) NDCT

Fig. 2. Result of Denoising for comparison. The display window for the abdomen image (top row) is set to $[-140, 260]$, and $[-1200, 600]$ for the chest image.

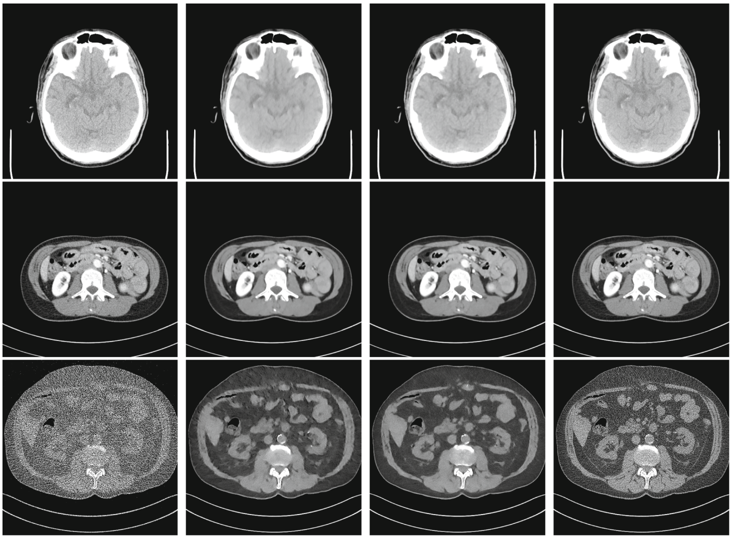
methods. It can be seen that the number of FLOPs of the dynamic weight modulation technique is not considerably higher than the baseline network M1, yet the improvement in performance is much appreciable.

In Fig. 2, we provide a visual comparison of the denoised output produced by different networks. Two sample images from datasets D1 and D2, corresponding to the abdomen and chest regions, respectively, are shown. The comparison shows that the proposed network M2 outperforms the baseline model M1 in terms of noise reduction and details preservation. For instance, in the denoised image of the abdomen region, the surface of the liver in M1 appears rough and splotchy due to noise, while in M2, the image is crisp, and noise suppression is adequate. Similarly, in the chest LDCT images, noticeable streaking artifacts near the breast region are present in the M1 output, and the boundaries of different organs like the heart and shoulder blade are not well-defined. In contrast, M2 produces crisp and definite boundaries, and streaking artifacts are significantly reduced. Moreover, M1 erases finer details like tiny blood vessels in the lung region, leading to compromised visibility, while M2 preserves small details much better than M1, resulting in output that is comparable with the original NDCT image.

Robustness Analysis: In this section, we evaluate the performance of existing denoising networks in a challenging scenario where the networks are trained to remove noise from a mixture of LDCT images taken from different anatomical regions with varying noise variances and patterns. We compared two networks in this analysis: M3, which is a baseline UNet model trained using a mixture of LDCT images, and M4, which is the proposed weight-modulated network, trained using same training data. Table 2 provides an objective comparison between these two methods. We found that joint training has a negative impact on the performance of the baseline network, M3, by a significant margin. Specifically, M3 yielded 0.88 dB lower PSNR than model M1 for head images, which

Table 2. Objective comparison among different methods. Objective metrics are reported by averaging the values for all the images present in the test set.

Model	Abdomen			Head			Chest		
M3	33.64	0.895	8.54	38.67	0.937	3.45	29.28	0.612	105.2
M4	34.17	0.921	7.45	39.70	0.964	2.12	29.69	0.689	89.21



(a) LDCT (b) M3 (c) M4 (d) NDCT

Fig. 3. Result of Denoising for comparison. The display window for the abdomen image is set to $[-140, 260]$, $[-175, 240]$ for the chest image, and $[-80, 100]$ for the head.

were trained using only head images. Similar observations were also noted for other anatomical regions like the abdomen and chest. The differences in noise characteristics among the different LDCT images make it difficult for a single model to denoise images efficiently from a mixture of anatomical regions. Furthermore, the class imbalance between small anatomical sites (e.g., head, knee, and prostate) and large anatomical locations (e.g., lung, abdomen) in a training set introduces a bias towards large anatomical sites, resulting in unacceptably lower performance for small anatomical sites. On the other hand, M4 showed robustness to these issues. Its performance was similar to M2 for all settings, and it achieved 0.69 dB higher PSNR than M3. Noise-conditioned weight modulation enables the network to adjust its weight based on the input images, allowing it to denoise every image with the same efficiency.

Figure 3 provides a visual comparison of the denoising performance of two methods on LDCT images from three anatomical regions. The adverse effects of joint training on images from different regions are apparent. Head LDCT images, which had the lowest noise, experienced a loss of structural and textural information in the denoising process by baseline M3. For example, the head lobes appeared distorted in the reconstructed image. Conversely, chest LDCT images, which were the noisiest, produced artefacts in the denoised image by M3, significantly altering the image's visual appearance. In contrast, M4 preserved all structural information and provided comparable noise reduction across all anatomical structures. CNN-based denoising networks act like a subtractive method, where the network learns to subtract the noise from the input signal by using a series of convolutional layers. A fixed set of subtracters is inefficient for removing noise from images with various noise levels. As a result, images with low noise are over smoothed and structural information is lost, whereas images with high noise generate residual noise and artefacts. In case of images containing a narrow range of noise levels, such as images from a single anatomical region, the above-mentioned limitation of naive CNN-based denoisers remains acceptable, but when a mixture of images with diverge noise levels is used in training and testing, it becomes problematic. The proposed noise conditioned weight modulation addresses this major limitation of CNN based denoising network, by designing an adjustable subtractor which is adjusted based on the input signal.

Figure 4 presents a two-dimensional projection of the learned embedding for all the test images using the TSNE transformation. The embedding has created three distinct clusters in the 2D feature space, each corresponding to images from one of three different anatomical regions. This observation validates our claim that the embedding learned by the anatomy encoder represents a meaningful representation of the input image. Notably, the noise level of low dose chest CT images differs significantly from those of the other two regions, resulting in a

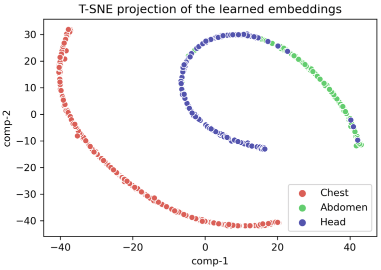


Fig. 4. 2 dimensional projection of learned embedding. The projection are learned using TSNE transformation.

Table 3. Objective comparison among different networks. Objective metrics are reported by averaging the values for all the images present in the test set of abdominal images taken with 1mm slice thickness.

Model	M5	M6	M7	M8
PSNR	22.23	22.55	22.80	22.96
SSIM	0.759	0.762	0.777	0.788
RMSE	32.13	30.13	29.37	29.14

separate cluster that is located at a slightly greater distance from the other two clusters.

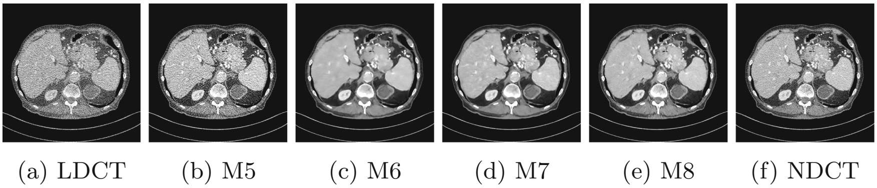

(a) LDCT (b) M5 (c) M6 (d) M7 (e) M8 (f) NDCT

Fig. 5. Result of Denoising for comparison. The display window for the abdomen image is set to $[-140, 260]$

Generalization Analysis: In this section, we evaluate the generalization ability of different networks on out-of-distribution test data using LDCT abdomen images taken with a 1mm slice thickness from dataset D1. We consider four networks for this analysis: 1) M5, the baseline UNet trained on LDCT abdomen images with a 3mm slice thickness from dataset D1, 2) M6, the baseline UNet trained on a mixture of LDCT images from all anatomical regions except the abdomen with a 1mm slice thickness, 3) M7, the proposed weight-modulated network trained on the same training set as M6, and 4) M8, the baseline UNet trained on LDCT abdomen images with a 1mm slice thickness. Objective comparisons among these networks are presented in Table 3. The results show that the performance of M5 and M6 is poor on this dataset, indicating their poor ability to generalize to unseen data. In contrast, M7 performs similarly to the supervised model M8. Next, we compared the denoising performance of different methods visually in Fig. 5. It can be seen that M5 completely failed to remove noise from these images despite the fact the M5 was trained using the abdominal image. Now the output of M6 is better than the M5 in terms of noise removal, but a lot of over-smoothness and loss of structural information can be seen, for example, the over-smooth texture of the liver and removal of blood vessels. M6 benefits from being trained on diverse LDCT images, which allows it to learn robust features applicable to a range of inputs and generalize well to new images. However, the CNN networks' limited ability to handle diverse noise levels results in M6 failing to preserve all the structural information in some cases. In contrast, M7 uses a large training set and dynamic convolution to preserve all structural information and remove noise effectively, comparable to the baseline model M8.

5 Conclusion

This study proposes a novel noise-conditioned feature modulation layer to address the limitations of convolutional denoising networks in handling variability in noise levels in low-dose computed tomography (LDCT) images. The proposed technique modulates the weight matrix of a convolutional layer according to the noise present in the input signal, creating a slightly modified neural network. Experimental results on two public benchmark datasets demonstrate that this dynamic convolutional layer significantly improves denoising performance, as well as robustness and generalization to unseen noise levels. The proposed method has the potential to enhance the accuracy and reliability of LDCT image analysis in various clinical applications.

References

1. Bera, S., Biswas, P.K.: Noise conscious training of non local neural network powered by self attentive spectral normalized Markovian patch GAN for low dose CT denoising. IEEE Trans. Med. Imaging **40**(12), 3663–3673 (2021). https://doi.org/10.1109/TMI.2021.3094525
2. Chen, H., et al.: Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
3. He, T., Shen, C., Van Den Hengel, A.: DyCo3D: robust instance segmentation of 3D point clouds through dynamic convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 354–363 (2021)
4. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
5. Kang, E., Chang, W., Yoo, J., Ye, J.C.: Deep convolutional framelet denosing for low-dose CT via wavelet residual network. IEEE Trans. Med. Imaging **37**(6), 1358–1369 (2018)
6. Kang, E., Min, J., Ye, J.C.: A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med. Phys. **44**(10), e360–e375 (2017)
7. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
8. Klein, B., Wolf, L., Afek, Y.: A dynamic convolutional layer for short range weather prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4840–4848 (2015)
9. McCollough, C.H., et al.: Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. Med. Phys. **44**(10), e339–e352 (2017)
10. Moen, T.R., et al.: Low-dose CT image and projection dataset. Med. Phys. **48**(2), 902–911 (2021)
11. Murphy, A., Bell, D., Rock, P., et al.: Noise (CT). Reference article, Radiopaedia.org (2023). https://doi.org/10.53347/rID-51832. Accessed 08 Mar 2023
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Sprawls, P.: AAPM tutorial. CT image detail and noise. Radiographics **12**(5), 1041–1046 (1992)
14. Yin, X., et al.: Domain progressive 3D residual convolution network to improve low-dose CT imaging. IEEE Trans. Med. Imaging **38**(12), 2903–2913 (2019)

15. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
16. Zhang, K., Zuo, W., Zhang, L.: FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)