Sparse-view CT reconstruction based on multi-level wavelet convolution neural network

Minjae Lee, Hyemi Kim, Hee-Joung Kim

A R T I C L E   I N F O
Keywords:
Sparse-view computed tomography
Reconstruction
Convolutional neural network
Multi-level wavelet

A B S T R A C T
Sparse-view computed tomography (CT) is a recent approach to reducing the radiation dose in patients and speeding up the data acquisition. Consequently, sparse-view CT has been of particular interest among researchers within the CT community. Advanced reconstruction algorithms for sparse-view CT, such as iterative algorithms with total-variation (TV), have been studied along with the problem of increasing computational burden and the blurring of artifacts in the reconstructed images. Studies on deep-learning-based approaches applying U-NET have recently achieved remarkable outcomes in various domains including low-dose CT. In this study, we propose a new method for sparse-view CT reconstruction based on a multi-level wavelet convolutional neural network (MWCNN). First, a filtered backprojection (FBP) was used to reconstruct a sparsely sampled sinogram from 60, 120, and 180 projections. Subsequently, the sparse-view data obtained from FBP were fed to a deep-learning network, i.e., the MWCNN. Our network architecture combines a wavelet transform and modified U-NE T without pooling. By replacing the pooling function with the wavelet transform, the receptive field is enlarged to improve the performance. We qualitatively and quantitatively evaluated the interpolation, iterative TV method, and standard U-NET in terms of a reduction in the streaking artifacts and a preservation of the anatomical structures. When compared with other methods, the proposed method showed the highest performance based on various evaluation parameters such as the structural similarity, root mean square error, and resolution. These results indicate that the MWCNN possesses a powerful potential for achieving a sparse-view CT reconstruction.

1. Introduction
Many researchers in the physical and engineering domains have attempted to solve the trade-off between image quality and radiation dose for X-ray or computed tomography (CT) imaging, both of which are widely used in clinical, industrial, and other applications [1–4]. The widespread application of CT in a clinical setting has directed attention toward the associated X-ray radiation dose, which may pose a lifetime risk of cancer [4]. Accordingly, researchers in this domain have aimed to achieve a radiation-dose-reduction technology in line with the principle of "as low as reasonably achievable" [5].
A representative strategy for lowering the X-ray radiation dose is to control the tube current or exposure time of the X-ray source. Although this can effectively reduce the radiation dose, it will increase the noise in the image. Another strategy is to reduce the number of projections for a given trajectory [6–8]. Sparse-view CT may not be appropriate for conventional multi-detector CTs because of the rapid and continuous acquisition of the projection view. However, many interesting studies have used an alternating kVp-switching technology and a dynamic beam blocker in spectral CT [9–12]. Moreover, sparse-view CT provides an opportunity to reduce the scan time in cone-beam CTs, such as dental, mammography, and C-arm CTs [13,14]. However, this approach results in significant streak artifacts owing to insufficient projection data. In this study, we focus on restoring the image quality from sparse-view CT with a limited scanning angle.

Various studies have been conducted to improve the quality of limited-angle CT [7,8,15,16]. In addition, several studies have attempted to reduce the artifacts of limited-angle CT by minimizing the energy or entropy of the images using techniques such as singular value decomposition and wavelet decomposition. However, this method is difficult to apply to general medical imaging because it requires prior knowledge of the image to be restored. In addition, considering the fact...
that compressed sensing techniques show an excellent noise reduction performance, studies using the total-variation (TV) minimization technique have been conducted [17]. However, studies using the TV technique have only been successful in removing local noise and not in removing linear artifacts that appear as large structures in limited-angle tomography. In particular, this technique has the disadvantage of requiring a significant computation time because it requires repetitive projections and back-projections during the iterative update procedure.

Many studies applying deep learning have recently been successful in various domains, including classification, segmentation, denoising, and super resolution [18–21]. For CT applications in the medical physics and imaging domains, a deep convolutional neural network (CNN) for low-dose CT was proposed, and it was shown that a CNN that uses a directional wavelet is effective in removing the noise associated with low-dose CT [21–23]. Unlike noise artifacts generated under the strategy of controlling the tube current or X-ray exposure of an X-ray source for low-dose CT, the streaking artifacts generated upon reducing the number of projections show globalized artifacts that are difficult to remove using conventional CNNs [25,26]. To improve the removal efficiency of streaking artifacts, residual learning networks applying U-Net have been proposed [27]. However, owing to the use of a CNN architecture with a small receptive field, such an approach is insufficient in eliminating globally distributed artifacts. Han et al. [28] demonstrated that a CNN architecture with a large receptive field is crucial to the effective removal of streaking artifacts.

Budda et al. [39] proposed a hybrid deep learning-shearlet framework for sparse CT. In their study, an iterative shearlet transform algorithm was used to reconstruct the visual singularities of an imaged object, and a U-Net architecture was used to predict invisibility. This method can provide better image quality than data-driven based reconstruction methods or pure models. Additionally, a different approach incorporating a physical forward model into the network was studied in [41]. However, these improved results are generally achieved at the cost of a longer computation time. Recent “noise to noise” training methods [40,42] have been shown to provide satisfactory image quality for CNNs with pairs of noise images in certain applications. However, the training of noise images may have certain limitations in sparse-view CT reconstruction [43].

In this study, we present a multi-level wavelet CNN (MWCNN) with a large receptive field for effectively removing streaking artifacts in low-dose CT. This network is based on the U-NET architecture [19]. In the contracting subnetwork, a discrete and inverse wavelet transform (IWT) is embedded to replace the pooling operation. This superosed function can obtain both the frequency and location information of the feature maps and effectively preserve the detailed textures [29,30]. Therefore, we propose a method for restoring a sparse-view CT using an MWCNN. Because the artifacts in a sparse-view CT appear along a clear direction according to a limited angle, the architecture was designed to adequately recognize and remove artifacts.

2. Materials and methods

2.1. Multi-level wavelet convolution neural network

The overall framework proposed in this study is depicted in Fig. 1. We converted a sufficient number of projections into sparsely sampled sinograms to reconstruct them using a filtered backprojection (FBP). As shown in Fig. 1(a), the reconstructed image is a well-known sparse-view CT reconstruction using FBP, which will result in numerous streaking artifacts in input X. Streaking artifacts are often generated in CT images because of missing data under certain scanning angles. We configured the convergence through our network using label Y, which sparsely subtracts sampled projections from full projections to remove the streaking artifacts from input X. A modified CNN-based architecture was previously used to remove streaking artifacts. However, the architecture placed in each subband pooling operator, such as max pooling or average pooling, may not be effective in removing streaking artifacts because the receptive field is reduced as the network deepens. In this study, our network offers an effective receptive field by replacing the pooling operator with a wavelet decomposition operator to remove globally distributed streaking artifacts.

Multi-level wavelet packet transforms (WPTs) are convolved using four filters \( f_{ll}, f_{lu}, f_{ul}, \) and \( f_{uh} \) in a two-dimensional discrete wavelet transform (DWT) [8]. The down-sampled results are the subband images \((x_1, x_2, x_3, \) and \( x_4 \)), which are defined as \( f_{ij} \otimes x \). The subband images \((x_1, x_2, x_3, \) and \( x_4 \)) are further decomposed using DWT, resulting in

![Fig. 1. (a) Framework of sparse-view CT reconstruction based on the proposed method. (b) Composition principle of down-sampling and up-sampling in four-level wavelet transform.](image-url)
four subband images, $DWTx_i = x_{i1}, x_{i2}, x_{i3},$, and $x_{i4}$. Because the DWT possesses biorthogonal characteristics, the original image, $x$, can be reconstructed using the IWT as follows: $x = IWT(x_1, x_2, x_3, x_4)$. The deeper the network structure is, the higher the level of the WPT, and the subband images, $x_i (i = 1, 2, 3, \text{ or } 4)$, are then repeatedly divided into the respective subband images. The basic principles of decomposition and reconstruction in an image using a WPT are shown in Fig. 1(b). A WPT is considered a special case of a fully convolutional network (FCN) except for the non-linearity layer. By substituting a WPT with predefined filters in each subband section where the pooling operator is deployed, both downsampling and upsampling is applied.

2.2. Network architecture

The designed MWCNN architecture is shown in Fig. 2. Each CNN block is a 5-layer FCN with the wavelet transform instead of pooling and takes all subband images as inputs. Each layer of the CNN block includes the convolution filter ($3 \times 3$ Conv), batch normalization (BN), and rectified linear unit (ReLU) function. The Conv, without BN and ReLU, of the last layer was used to predict the residual image. By contrast, different CNNs are placed within the low- and high-frequency bands of deep convolutional framelets [21,24]. Note that the subband images after DWT are still dependent and not knowing such dependencies can be detrimental to restore performance. The designed network is a modified U-NET. A conventional U-NET generally uses max-pooling and an up-convolution, whereas MWCNN uses a DWT and an IWT. In the case of MWCNN, downsampling results in an increase in the feature map channels. Excluding the first block, the other blocks are placed to reduce the feature map channels for dense representation. For conventional U-NET, the downsampling does not affect the feature map channel and uses subsequent convolutions to increase the channel. For the designed MWCNN, an element-by-element summation is used to combine the feature maps in contracting and expanding subnetworks. The concatenation is the same as that used in the conventional U-NET. Our final network contains 32 layers. The wavelet operation used in this study adopted a Haar wavelet as the default.

2.3. Network training

In this study, the input size of the sparse-view CT is $512 \times 512$. The MWCNN was trained using the ADAM algorithm with $\alpha = 0.01, \beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$ for optimization. The learning rate exponentially decayed from $10^{-4}$ to $10^{-6}$ in 100 epochs. The main advantage of the ADAM algorithm is that the step-size is not affected by the rescaling of the gradient. Even if the gradient is large, the step-size is bound such that any objective function can be used to steadily descend for optimization. The ADAM algorithm can adapt the step-size by referring to the previous gradient size. The network was implemented using the MatConvNet toolbox with cuDNN 6.0 in MATLAB 2019a. We also used an Nvidia GTX 1080 Ti GPU.

2.4. Data acquisition and data set

The CT system (Ingenuity, Philips) was used to acquire phantom data as shown in Fig. 3. The tube voltage and tube current are 120 kVp and 80 mAs, respectively, and measured dose is 5.2 mGy. In addition, the voxel size used for FBP reconstruction is $0.66 \times 0.66 \times 2.5 \text{ mm}^3$. The resolution obtained is 1.505 pixel/mm. The exposure time per projection was measured to be 0.494 s. A LUNGMAN (Kyoto Kagaku Co., Ltd, Kyoto, Japan) was applied to examine the clinical applicability of our methods. Compared with that of real patients, there is no lung parenchyma in the phantom. The subsolid nodules were inserted as shown in Fig. 3. Nodule phantoms with various diameters were inserted at $\approx 800 \text{ HU}$ to $+100 \text{ HU}$ in addition to those shown in Fig. 3. The image objects were also acquired using a Catphan® (The phantom Laboratory, Salem, USA) to measure the quality evaluation, such as a modulation transfer function (MTF).

![Fig. 2. MWCNN architecture for the restoration of sparse-view CT.](image)

Table 1 shows the data configuration. We augmented the training data through simple processing such as image rotation and flipping to prevent overfitting of learning data. For the training data of the chest phantom, 1889 image slices with a pixel resolution of $512 \times 512$ were applied. The test data were composed of 300 image slices with the nodules at different positions. The primarily result data selected were inserted lung nodule slices; the Catphan phantom with 280 data was

<table>
<thead>
<tr>
<th>Data type</th>
<th>Volume size</th>
<th>Training</th>
<th>Testing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest phantom – 1</td>
<td>$512 \times 512 \times 120$</td>
<td>120</td>
<td>–</td>
<td>420</td>
</tr>
<tr>
<td>Chest phantom – 2</td>
<td>$512 \times 512 \times 300$</td>
<td>300</td>
<td>–</td>
<td>600</td>
</tr>
<tr>
<td>Chest phantom – 3</td>
<td>$512 \times 512 \times 300$</td>
<td>–</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>Catphan phantom – 1</td>
<td>$512 \times 512 \times 80$</td>
<td>80</td>
<td>–</td>
<td>280</td>
</tr>
<tr>
<td>Catphan phantom – 2</td>
<td>$512 \times 512 \times 200$</td>
<td>–</td>
<td>200</td>
<td>–</td>
</tr>
</tbody>
</table>
applied to analyze the image sharpness and MTF only using the training data and not the testing data. We simply removed the number of projections from the full projections using FBP and used them as input X. Label data Y were used as residual images between 60, 120, and 180 projections and full projection scans.

2.5. Quantitative evaluation

The peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and normalized mean square error (NMSE) values were used, which are defined as follows:

$$PSNR = 20 \cdot \log_{10}\left( \frac{MN}{\|x - \hat{x}\|_2} \right)$$

(1)

$$NMSE = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - \hat{x}(i,j))^2}{\sum_{i=1}^{M} \sum_{j=1}^{N} |x(i,j)|^2}$$

(2)

$$SSIM = \frac{(2\mu_x \mu_{\hat{x}} + \mu_x^2 + C_1)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)}$$

(3)

where $x$ and $\hat{x}$ denote the reconstructed image and ground truth, respectively. The terms, $M$ and $N$, denote the numbers of pixels for the row and column, respectively. In addition, $\mu_x$ and $\mu_{\hat{x}}$ denote the local averages of $x$ and $\hat{x}$, respectively. The term, $\sigma_x^2$, denotes the covariance of $x$ and $\hat{x}$. The normalization constants, i.e., $C_1$ and $C_2$, were used to avoid instability in image areas where the local mean or standard deviation was close to zero. Notably, $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$. In addition, $k_1$ and $k_2$ are constants by default, and their values are 0.01 and 0.03, respectively. The term, $L$, denotes the dynamic range of the pixel intensities. The higher the PSNR value is, and the closer the NMSE value reaches zero, the higher the image quality of the compared image. The SSIM ranges between zero and 1, where 1 is reached when both value reaches zero, the higher the image quality of the compared image.

In this study, we adopted additional evaluation methods, namely, multi-scale structural similarity (MS-SSIM) and feature similarity (FSIM), for a more accurate quantitative evaluation. The SSIM evaluation method used in the past has often been modified into MS-SSIM, because the former presents a critical error in the evaluation according to the image scale. The multi-scale version of SSIM, MS-SSIM, is expressed [9] as follows:

$$MS - SSIM(x,y) = [l(x,y)]^{M^2} \times \prod_{j=1}^{M} [c(x,y)]^{b_j} \times [s(x,y)]^{g_j}$$

(4)

$$l(x,y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

(5)

$$c(x,y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

(6)

$$s(x,y) = \frac{2\sigma_x \sigma_y + C_3}{\sigma_x^2 + C_3}$$

(7)

where the luminance, contrast and structure are $l(x,y)$, $c(x,y)$, and $s(x,y)$. Contrast and structural factors are weighted for each scale, and the luminance is calculated only for full-resolution images. The constants are given by $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$, and $C_3 = C_2/2$. We set $\alpha M = \beta_j = \gamma_j$ for the full j-th scale. The cross-scale settings are normalized by $\sum_{j=1}^{M} \gamma_j = 1$.

In addition, more advanced evaluation methods have been proposed, such as a gradient-based SSIM, the use of a quality index based on local variance (QILV) using the same formula, the gradient similarity, and a three-component SSIM. Among these methods, an interesting modification of this general idea has been proposed using the Riesz transform, resulting in the Riesz FSIM metric, which is the basis for the widely accepted FSIM metric [10,11] defined as follows:

$$FSIM = \frac{\sum_{u,v} S(u,v) \times PC_{max}(u,v)}{\sum_{u,v} PC_{max}(u,v)}$$

(8)

with phase congruency (PC) and gradient magnitude (G) factors calculated using the SSIM as in the following:

$$S(u,v) = \frac{2 \times PC_1(u,v) \times PC_2(u,v) + T_{PC}}{PC_1(u,v) + PC_2(u,v) + T_{PC}} \times \frac{2 \times G_1^2(u,v) \times G_2^2(u,v) + T_{G}}{G_1^2(u,v) + G_2^2(u,v) + T_{G}}$$

(9)

where $PC_{max}$ denotes the higher of two local values of PC from the reference and assessed image, $(u,v)$ denotes the position of the sliding window in images 1 and 2, and $T$ is used to prevent dividing by zero. We set $T_{PC} = 0.85$ and $T_{G} = 160$ in Eq. (9). The gradient values can be obtained using the convolution filter (Scharf mask). The MS-SSIM and FSIM ranges between zero and 1, where 1 is reached when both images compared are identical.

3. Results

Herein, we provide objective indicators using more accurate and diverse evaluation methods, as shown in Table 2. We listed the PSNR, NMSE, SSIM, MS-SSIM, and FSIM values of the different methods and

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>NMSE</th>
<th>SSIM</th>
<th>MS-SSIM</th>
<th>FSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBP - 60</td>
<td>21.8 ± 0.54</td>
<td>0.14 ± 0.03</td>
<td>0.31 ± 0.02</td>
<td>0.31 ± 0.02</td>
<td>0.93 ± 0.004</td>
</tr>
<tr>
<td>Interpolation - 60</td>
<td>25.0 ± 0.57</td>
<td>0.06 ± 0.01</td>
<td>0.40 ± 0.03</td>
<td>0.42 ± 0.03</td>
<td>0.94 ± 0.003</td>
</tr>
<tr>
<td>U-NE T-60</td>
<td>27.8 ± 0.58</td>
<td>0.07 ± 0.01</td>
<td>0.72 ± 0.03</td>
<td>0.70 ± 0.04</td>
<td>0.96 ± 0.005</td>
</tr>
<tr>
<td>Proposed - 60</td>
<td>38.0 ± 0.80</td>
<td>0.002 ± 0.0004</td>
<td>0.92 ± 0.01</td>
<td>0.94 ± 0.01</td>
<td>0.97 ± 0.001</td>
</tr>
<tr>
<td>FBP - 120</td>
<td>40.1 ± 1.47</td>
<td>0.002 ± 0.0007</td>
<td>0.95 ± 0.01</td>
<td>0.96 ± 0.01</td>
<td>0.98 ± 0.004</td>
</tr>
<tr>
<td>Interpolation - 120</td>
<td>26.9 ± 0.47</td>
<td>0.04 ± 0.009</td>
<td>0.44 ± 0.02</td>
<td>0.53 ± 0.02</td>
<td>0.95 ± 0.005</td>
</tr>
<tr>
<td>U-NE T-120</td>
<td>30.0 ± 0.52</td>
<td>0.02 ± 0.004</td>
<td>0.60 ± 0.02</td>
<td>0.68 ± 0.02</td>
<td>0.97 ± 0.005</td>
</tr>
<tr>
<td>Proposed - 120</td>
<td>35.5 ± 3.33</td>
<td>0.08 ± 0.109</td>
<td>0.87 ± 0.01</td>
<td>0.88 ± 0.01</td>
<td>0.98 ± 0.002</td>
</tr>
<tr>
<td>FBP - 180</td>
<td>42.3 ± 0.88</td>
<td>0.0012 ± 0.0002</td>
<td>0.95 ± 0.01</td>
<td>0.96 ± 0.01</td>
<td>0.99 ± 0.009</td>
</tr>
<tr>
<td>Proposed - 180</td>
<td>43.0 ± 1.56</td>
<td>0.0011 ± 0.0005</td>
<td>0.97 ± 0.01</td>
<td>0.97 ± 0.01</td>
<td>0.99 ± 0.003</td>
</tr>
<tr>
<td>FBP - 240</td>
<td>30.8 ± 0.52</td>
<td>0.01 ± 0.003</td>
<td>0.63 ± 0.03</td>
<td>0.71 ± 0.02</td>
<td>0.96 ± 0.005</td>
</tr>
<tr>
<td>Interpolation - 240</td>
<td>33.1 ± 0.51</td>
<td>0.01 ± 0.002</td>
<td>0.76 ± 0.02</td>
<td>0.82 ± 0.01</td>
<td>0.97 ± 0.004</td>
</tr>
<tr>
<td>U-NE T-240</td>
<td>35.3 ± 0.78</td>
<td>0.006 ± 0.001</td>
<td>0.94 ± 0.01</td>
<td>0.95 ± 0.01</td>
<td>0.98 ± 0.004</td>
</tr>
<tr>
<td>Proposed - 240</td>
<td>44.1 ± 1.05</td>
<td>0.0007 ± 0.0002</td>
<td>0.97 ± 0.007</td>
<td>0.97 ± 0.007</td>
<td>0.99 ± 0.005</td>
</tr>
<tr>
<td>FBP - 300</td>
<td>44.1 ± 0.77</td>
<td>0.0009 ± 0.0002</td>
<td>0.98 ± 0.003</td>
<td>0.98 ± 0.003</td>
<td>0.99 ± 0.001</td>
</tr>
</tbody>
</table>

Table 2

Quantitative evaluation of different methods for 60, 120, and 180 projections. Evaluations were performed with 120 different slices, and the mean and standard deviation were calculated for a total of 120 different slices.
the proposed approach upon applying these methods to a sparse-view CT from limited projection views (i.e., 60, 120, and 180 projections). All methods provided superior quantitative indicators compared with FBP, obtained using limited projections. Among all methods considered, the proposed approach produced the best quantitative values, followed by the standard U-NET. If parameters such as the size of the window and the gradient magnitude factors for MS-SSIM and FSIM are adjusted, the quantitative values may change. However, the performance of the proposed network was not significantly different from those of the other methods and showed the best values for all other evaluation indicators.

In addition, a visual inspection showed the advantages of both U-NET and the proposed method. Specifically, in Fig. 4, we compare the reconstruction results obtained using the interpolation, projection onto the convex sets (POCS)-TV, U-NET, and MWCNN methods with the ground-truth (original) image. As shown in Interpolation and POCS-TV in Fig. 4(a) and (a1), although the streak artifacts were reduced, as indicated in Table 2, the quantitative values were relatively reduced, and thus were not completely removed. The performance of the interpolation and POCS-TV significantly decreased as the number of projections decreased, and POCS-TV could not completely preserve the information because of the blurring of the TV. However, U-NET and MWCNN efficiently removed the streak artifacts and preserved the information of the anatomical structure in 60, 120, and 180 of the views.

The results obtained through interpolation, POCS-TV, U-NET, and the proposed method for 60 projections are shown in Fig. 5. We inserted lung nodules with −20 and −630 HU in the 52nd reconstructed image. The residual is the difference image between the ground-truth and CT images reconstructed using the different comparison methods and the proposed method. Notably, the proposed method showed the smallest residual, and the results for the interpolation, POCS-TV, and U-NET still include visible streaking artifacts and other structures. In Fig. 5, the darker the reconstruction result, the smaller the error. POCS-TV may achieve different results depending on the adjustment of the parameters; however, as the number of projections decreases, such as from 60 and

![Fig. 4. Reconstruction results from interpolation, POCS-TV, U-NET, and MWCNN for various sparse-view reconstructions.](image-url)
For greater detail, Fig. 6 shows the enlarged areas marked by the yellow, orange, and red boxes in Fig. 5. The enlarged images in Fig. 5(a) and (c) depict lung nodules with -20 and -630 HU, respectively, and Fig. 6(b) depicts the lung vessels. From Fig. 6(a), it is clear that the proposed method shows the least structural loss in the lung nodule with -20 HU. In addition, U-NET demonstrates a good performance. We
provide the enlarged image in Fig. 6(b) to better analyze the detailed structure of the lung. However, U-NET in Fig. 6(b) shows a loss of detail in the structure. In addition, the lung nodule with a relatively low HU showed structural loss, as depicted in Fig. 6(c). Comparatively, the proposed method shows a satisfactory image restoration in all three enlarged images.

In Fig. 7, we show an image used to measure and evaluate a high-resolution module, called CTP528, in a Catphan phantom. This slice section has a high-resolution test gauge with 1–21 line pairs per centimeter. From the reconstruction results obtained using the different methods and the proposed approach, the beads for the 5–7 line pairs per centimeter, as shown in the red box in Fig. 7, were enlarged, and the profile was analyzed, as depicted in Fig. 8. For a detailed performance evaluation, the residual images are also shown in the bottom images in Fig. 7. From the profile results of Fig. 8, it is clear that the proposed method matches exactly the reference at 5 line pairs per centimeter, and the profiles of 6 and 7 line pairs per centimeter were almost identical. By contrast, the resolution of the results obtained from the interpolation and POCS-TV was significantly reduced, as depicted in Figs. 7 and 8. Next, it was demonstrated that U-NET satisfactorily reduced the streaking artifacts and restored the overall structure, as shown in Figs. 4 and 5. However, the restoration of the detailed structure for a high resolution is not as satisfactory as that obtained using the proposed method, as depicted in Figs. 6 and 7.

In Fig. 9, we measured MTF for POCS-TV, U-NET, and MWCNN using the Catphan phantom’s bead point source. The sparse-view and interpolation methods are excluded because measurements were difficult to achieve owing to severe streak artifacts. The MTF performance is the best for the proposed method and was measured in order of U-NET and POCS-TV.

Comparisons of the total number of parameters, operations, and
training time are shown in Table 3. The total number of parameters of MWCNN is approximately 31.2 M, which is similar to that of U-NET. The numbers of down- and up-scaling operations for both U-NET and MWCNN are four. The training time for MWCNN was 0.46 s/slice and 24.1 h/100 epochs.

The convergence plots for loss function, mean square error (MSE), PSNR, and SSIM were measured for each epoch, as shown in Fig. 10. The convergence plots show the advantages of the proposed method over U-NET. The proposed network exhibited the fastest convergence, and the final performance was better.

4. Discussion

In recent years, a CNN applied in the medical domain has been studied as a mapping technique for converting degraded images into latent clean images. U-NET, a highly researched approach, is a representative of a CNN and has been used in various domains such as segmentation, classification, and denoising. The problem of a pooling layer, which is present in a CNN, has been considered for some time. For example, the input and output images must have the same size; therefore, using an FCN without a pooling layer is a typical strategy. A network without pooling layers can generally aid in improving the performance by considering more spatial context because the receptive field size can be enlarged through a larger filter size or deeper network depth. However, a network structure without pooling can result in an increase in the computational cost. Accordingly, the studies in [31] and [32], which applied dilated filtering without sacrificing the computational cost, have been presented; however, it may be accompanied by a gridding effect because the receptive field only considers sparse sampling for an input with checkerboard patterns [33]. Therefore, we should carefully enlarge the receptive field while avoiding the potential trade-off relationship between the computational cost and performance.

Table 3
Detailed comparisons of the tested CNNs.

<table>
<thead>
<tr>
<th></th>
<th>Total number of parameters</th>
<th>Number of down-sampling / up-sampling operations</th>
<th>Time per slice (sec)</th>
<th>Total time for training 100 epochs (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-NET</td>
<td>31030593</td>
<td>Max-pooling (2 × 2): 4 / Up-convolution (2 × 2): 4</td>
<td>0.26</td>
<td>13.6</td>
</tr>
<tr>
<td>MWCNN</td>
<td>31191552</td>
<td>DWT: 4 / IWT: 4</td>
<td>0.46</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Although the fast super-resolution convolutional neural network (FSRCNN) [34] has a relatively short run time with a large receptive field, it achieves a lower PSNR value than those of ResNet [35], DnCNN [20,36], and VDSR [37].

In this study, we proposed an MWCNN with an enlarged receptive field for improving the image restoration and enabling the efficient removal of streaking artifacts from a sparse-view CT. Our network is based on the U-NET [38] architecture. In the connecting subnetwork, the pooling operation is replaced by a DWT and an IWT. Because a DWT can be reversed, all information can be maintained using a downsampling scheme. In addition, a DWT can capture both the location information and frequency from the feature map [29,30], thereby maintaining a detailed texture. An IWT applied in an extended subnet is used to upsample low-resolution feature maps into high-resolution maps. To combine the feature maps from a DWT and an IWT, we applied an element-wise summation, which can enrich the feature representation and reduce the computational burden. Our dilated filtering, which is based on a wavelet transform, is a general and effective way to enlarge the receptive field. The experiment on artifact removal verified the effectiveness of the proposed method. However, the
proposed method may be slightly slower in terms of the runtime than U-NET, but it showed a higher quantitative evaluation with a larger receptive field.

From the comparison results shown in Figs. 6–9, it is evident that U-NET produces blurred edge images, whereas the proposed network enhances the high-frequency characteristics of the images. Despite achieving a better subjective quality, the standard U-NET offers better PSNR values for the entire image area than those of our network, as indicated in 180 views of Table 2, because of the larger noise amplification factor, which may be dominant in the background and boundary of the image, such as in a patient’s bed. As another reason for this case, the duplication of a low frequency may have contributed to the better PSNR value. However, compared with different methods that include standard U-NET, our network provides not only a better performance in other objective quality evaluations it also improves the visual quality.

The TV minimization hyperparameter for normalization was selected through trial and error to obtain the optimal balance between the resolution and mean square error, thereby resulting in a value of $2 \times 10^{-3}$.

The iterative method employed POCS to enforce the image constraints and steepest descent with an adaptive step-size to reduce the TV image [8]. The parameters of POCS-TV used in this study are provided in Table 4.

As the number of projection views decreased, the number of iterations gradually increased to optimize the image quality. The 60, 120, and 180 views for convergence are appropriate for 200, 150, and 100 iterations in our study, respectively. However, for POCS-TV, depending on the number of iterations, parameters, and object characteristics, the computational time may act as an important element in the reconstruction processing of the medical field. The computational time for the POCS-TV is approximately 60–100 s/slice with a GPU, whereas that for the proposed method with a GPU is approximately 0.4–0.5 s/slice. U-NET also showed a short reconstruction time of approximately 0.2–0.3 s/slice. This implies that the deep-learning method is significantly more powerful in terms of the computational time than the TV approach, resulting in a considerably better reconstruction performance. The MWCNN can combine DWT with a CNN to enlarge the receptive field without losing information, and it can be used by embedding a CNN with pooling and a DWT. Moreover, the embedded DWT can jointly adjust the dynamic range of the subbands using the CNN block, and can be treated as a predefined parameter to facilitate network learning. Therefore, the proposed network is powerful for modeling the interband dependency using all subbands as an input. Our network can be expanded to a higher level of wavelet decomposition. Because a higher level results in unavoidable consequences for deeper networks and a heavier computational burden, the appropriate levels are required to balance the efficiency and performance.

### Table 4
Parameter settings of POCS-TV for 60, 120, and 180 projections.

<table>
<thead>
<tr>
<th>Parameter settings for POCS-TV</th>
<th>180 views</th>
<th>120 views</th>
<th>60 views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of iterations</td>
<td>100</td>
<td>150</td>
<td>200</td>
</tr>
<tr>
<td>TV hyper-parameters</td>
<td>$2 \times 10^{-3}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of TV iterations</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Reduction of the TV hyper-parameter every iteration</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum allowed image/TV update ratio</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 11. Reconstruction results using U-NET and MWCNN for various sparse-view reconstructions. Lung nodules with different positions and various HU values were inserted for the testing data.](image-url)
5. Conclusion

In this study, we compared the performance of the proposed method with those of the linear interpolation, POCS-TV, and U-NET architecture. The results obtained using the existing methods have often included the blurring of the image and false restorations. Accordingly, we proposed an MWCNN to overcome the existing limitations and proved that our network architecture with a larger receptive field was superior in terms of the image restoration of globally distributed streaking artifacts than the other methods, such as the existing U-NET for a sparse-view CT reconstruction. Consequently, the proposed network produced significant results and might be an excellent alternative to low-dose CT applications.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1F1A1075741). Following are results of a study on the “Leaders in Industry-university Cooperation +” Project, supported by the Ministry of Education and National Research Foundation of Korea.

Appendix

We inserted lung nodules in different positions for the testing data. The testing data contained 300 slices, and the result was slices with lung nodules inserted as shown in Fig. 11. Nodules with relatively high HU values were well represented despite the small number of projections. However, nodules with small HU values lost information with 180 and 120 projections. Our network shows higher performance than U-NET.

To better illustrate the effectiveness of streak artifact removal by the proposed method, Fig. 12 shows the difference images between the original CT image and the reconstructed CT images achieved from the different methods: sparse FBP (60, 120, and 180 projections), interpolation, POCS-TV, U-Net, and the proposed approach. The proposed method achieves the smallest difference, whereas the other methods still include visible streaking artifacts and other structures. In Fig. 11, the darker the color, the smaller the error.

References


