Computed tomography image reconstruction using stacked U-Net

Computed Tomography Image Reconstruction using Stacked U-Net

Satoru Mizusawa[a,*], Yuichi Sei[a], Ryohei Orihara[a] and Akihiko Ohsuga[a]

[a]*The University of Electro-Communications 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan*

ARTICLE INFO

ABSTRACT

Since the development of deep learning methods, many researchers have focused on image quality improvement using convolutional neural networks. They proved its effectivity in noise reduction, single-image super-resolution, and segmentation. In this study, we apply stacked U-Net, a deep learning method, for X-ray computed tomography image reconstruction to generate high-quality images in a short time with a small number of projections. It is not easy to create highly accurate models because medical images have few training images due to patients' privacy issues. Thus, we utilize various images from the ImageNet, a widely known visual database. Results show that a cross-sectional image with a peak signal-to-noise ratio of 27.93 db and a structural similarity of 0.886 is recovered for a 512×512 image using 360-degree rotation, 512 detectors, and 64 projections, with a processing time of 0.11 s on the GPU. Therefore, the proposed method has a shorter reconstruction time and better image quality than the existing methods.

1. Introduction

In this study, we propose a method for X-ray computed tomography (CT) image reconstruction that is faster than the existing methods. X-ray CT image reconstruction is the process of reconstructing a cross-sectional image from a sinogram image captured by proton irradiation of X-ray detectors at various angles into the object. Projection is the process of irradiating objects with X rays produced by a generator and capturing them with a detector on the opposite side. An object is projected at various angles, and a sequence of images obtained at each angle is combined for composing a sinogram. The number of detectors and the projection angle can characterize the sinogram image's coordinate. The intensity of X rays captured by the detector may be attenuated to a particular extent depending on the object's nature. The particular extent of attenuation depends on the specific characteristics of an object in an X-ray's path. Suppose a bone or another obstacle is in the path; in that case, the intensity of X ray registered by the detector will deteriorate accordingly. The detector measures the overall result of X-ray absorption for an object in the X-ray's path. Based on the sinogram image obtained from this procedure, a cross-sectional image representing a distribution of a target object is derived through X-ray CT image reconstruction.

There are two main methods for X-ray CT image reconstruction: direct method and iterative reconstruction method. The direct method corresponds to the filtered back-projection (FBP) method (Kak and Slaney, 2001), whereas the iterative reconstruction method is based on the simultaneous algebraic reconstruction technique (SART) (Andersen, 1984, 1989) and maximum-likelihood expectation-maximization algorithm (Dempster et al., 1977). The direct method and iterative reconstruction method are different in terms of computational cost, noise, and artifacts.

The direct method has a low computational cost; however, if there is an insufficient number of projections in reconstruction, the resulting image's noise and artifacts will increase. The FBP method underlying the direct method is an analytical method combining filter calculations and Fourier inverse transformations. It is utilized in various CT devices. However, when the FBB method is applied to the results of a projection below the Nyquist frequency, the reconstructed cross-sectional image is markedly artifactual and noisy. For example, an obese patient's CT image is affected by high noise. The calcifications or stents cause blooming artifacts, and the metallic implants or bone structure cause streak artifacts (Geyer et al., 2015).

The iterative reconstruction method can achieve low noise and artifacts results, but its computational cost is relatively high. The iterative reconstruction method is based on an algebraic method, implying iteratively modifying an initial value in approaching the problem's solution. This method can reduce artifacts and noise even in the cases below

the Nyquist frequency because it considers the cross-sectional image's sparseness. However, this method needs to perform the iterative operation, resulting in high computational cost.

In the medical field, the number of projections should be reduced as much as possible to limit X rays' influence on a patient. Therefore, to address this problem, the possibility of developing an iterative approximation method has been investigated in recent years. Studies were conducted to reduce the computational cost and time of reconstruction to adapt for clinical use (Hudson and Larkin, 1994; Kim et al., 2015). The time required for reconstruction by the iterative reconstruction method ranges from 10 to 90 min. This cannot be used for emergent indications. However, noise reduction can be applied to patients considered challenging to treat, such as obese patients. By further reducing the radiation dose, we can expect to apply it to screening tests and other applications, such as lung cancer, colon cancer, and pediatric imaging (Geyer et al., 2015). Screening for COVID-19 can also reduce the patients' burden.

The advancement of deep learning methods has induced a considerable number of research works focusing on image quality improvement using convolutional neural networks (CNNs) (LeCun et al., 1989; Krizhevsky et al., 2017) to noise reduction (Lefkimmiatis, 2017), single-image super-resolution (Lai et al., 2017), and segmentation (Ronneberger et al., 2015).

In reconstruction, deep learning methods can recover cross-sectional images from a sinogram image. Yang et al. (2016) proposed a method that can achieve better results than the existing methods by replacing each step of sequential reconstruction in magnetic resonance imaging (MRI) using the alternating direction method of multipliers (Boyd et al., 2010) based on a deep learning method. Yang et al. (2018) applied generic adversarial network (GAN) (Mirza and Osindero, 2014) for MRI reconstruction. They demonstrated that it was possible to recover the data in a short processing time of 5 ms by training the model on a dataset, including 16,095 images. Zhu et al. (2019) proposed a model that combines a network to extract disease ROIs and a GAN and performs super-resolution on MRI images. Using ROIs can obtain faster convergence of super-resolution images of the lesion. The output of high-resolution images from the trained model was surprisingly fast, at 0.22–0.33 ms per image. They also proposed a mean opinion score, a subjective evaluation criterion by medical professionals. Using VGG loss as the evaluation function and adding low-resolution output to the generator output, they obtained results that surpassed the existing methods (Zhu, Jin; Guang, Yang; Pedro, Ferreira; Andrew, Scott, Sonia, Nielles-Vallespin; Jennifer, Keegan; Dudley and Pietro, Lio; David, 2019; Zhu et al., 2019; Yu et al., 2017). To reconstruct CT images, Jin et al. (2017) combined the FBP method with U-Net (Ronneberger et al., 2015) and residual learning (Kim et al., 2016; He et al., 2016). Similarly, Zhang et al. (2018) combined DenseNet with FBP (DD-Net). This method can reduce the amount of noise and artifacts generated by the FBP method, achieving better results than the existing methods.

U-Net is a CNN model with the same size of image input and output. The high-resolution features in this method are passed to the later stage by skip connection. Then, U-Net extracts the low-resolution features by repeatedly performing convolutional operations on high-resolution images and reducing the convolved images. The extracted low-resolution features are iteratively enhanced to produce a high-resolution image finally.

Several related studies have improved the U-Net's initial version. Sevastopolsky et al. (2019) showed that stacked U-Net and Res-U-Net in 15 stacks is better than U-Net alone in the segmentation task. Shah et al. (2018) let stacked U-Net to learn the classification task, showing that they could get results comparable with the existing methods. They further applied transfer learning to the model, obtaining results comparable to the existing methods in the segmentation task. Furthermore, Jegou et al. (2017) adapted DenseBlok as a layer structure, and they stacked the hourglass network similar to U-Net (Newell et al., 2016).

In this study, we aim to develop a reconstruction method for CT images by requiring a small number of projections, providing consistent image quality and faster processing speed compared with the existing iterative reconstruction methods. In the proposed method, we use deep learning as technology and employ stacked U-Net as a model architecture.

Generally, medical images are not easy to collect for training datasets because of patient privacy protection. Obtaining an actual cross-sectional image by CT is impossible because the patient's body cannot be invaded. The only possible option is using the results of cross-sectional image reconstruction through the existing methods.

Non-sparse sinograms reconstructed by FBP are used as the training data in FBP U-Net. However, during FBP reconstruction, noise caused by quantization errors affects the training data. Thus, the learned model is employed to reproduce the FBP method's results that are different from the actual cross-sectional image.

Here, we use ImageNet (Krizhevsky et al., 2017) for training. ImageNet is a dataset of natural images not captured by a medical device. From this dataset, we use 28,463 images as a training dataset. The model is trained using 300 iterations for 90 h.

From a simulation with 360-degree rotation, 512 detectors, and 64 projections, a cross-sectional image with a peak signal-to-noise ratio (PSNR) of 27.93 db and a structural similarity (SSIM) (Wang et al., 2004) of 0.886 is reconstructed in 2.35 s on the CPU and 0.11 s on the GPU.

2. Problem definition

CT image's projection is defined as follows:

$$y = Ax + b$$

where $y \in \mathbb{R}^{N_y}$ is the sinogram image, $A = a_{ij} \in \mathbb{R}^{N_y \times N_x}$ is a projection matrix that expresses the projection of X rays radiating around an object, $x \in \mathbb{R}^{N_x}$ denotes the cross-sectional image to be restored, and $b \in \mathbb{R}^{N_y}$ refers to noise.

Calculating A^{-1} and removing the effect of noise b are required in the reconstructing cross-sectional image x:

$$x = A^{-1}y - b$$

where A generally corresponds to a singular matrix, which is not easy to calculate explicitly.

Therefore, reconstruction of x can be expressed as an optimization problem formulated according to the following equation:

$$\arg\min ||y - Ax||^2 + \lambda norm(A)$$

where $\lambda > 0$. In this expression, the iterative reconstruction method sets the initial value and solves it iteratively. Here, we utilize a neural net W with y as the teacher input and x as the correct output and reconstruct x from y by approximating A^{-1}.

$$x = W y$$

3. Proposed Method

3.1. Structure of stacked U-Net

Figure 1 shows the structure of the stacked U-Net used in this study. The basic structure is the same as that of U-Net but connected using six stacks (Figure 2).
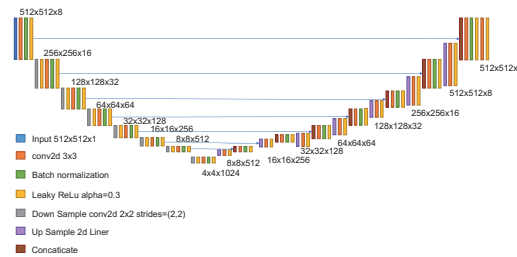

Figure 1: Structure of the proposed stacked U-Net

The original U-Net performs semantic segmentation of images; thus, the underlying task differs from the reconstruction. We adopt this model because of its same characteristics that generate images from images.

The definition of stacked U-Net is provided in the literature (Sevastopolsky et al., 2019; Shah et al., 2018). However, semantic segmentation has been considered as a target task. For classification purposes, the model's final output is the n-class output.

In this study, the final output is set to a single channel; thus, the output result is represented as an image; the intermediate output result is the LekyReLU function result; the final output result is normalized between 0 and 1 by adopting the sigmoid function.

Shah et al. (2018) discussed that stacked U-Net requires an excessive number of parameters, resulting in increased memory size that hinders training on a realistic GPU. Therefore, they adjusted the 3×3 convolution from 2 to 1 and

further adapted the scaling process to the 3 × 3 stride 3 convolution by setting the number of layers to three. In this study, we also reduce the memory size by decreasing the 3 × 3 convolution per layer from 2 to 1 compared with U-Net.

Shah et al. (2018) inputted the images into U-Net by reducing the preprocessing size, showing that three layers are sufficient. However, we set the layers to five layers because detailed features could not be extracted if the layers are made shallow. Various layer functions are also considered to improve the estimation accuracy.

We also modify the downscaling process from max pooling to 2 × 2 stride 2 convolution. We change the upscaling process from deconvolution to linear scaling and the activation function from the rectified linear unit (ReLU) to LekyReLU. We also apply batch normalization to the model.

4. Experimental Results

4.1. Selecting layer functions in 32 × 32 reconstruction

The first experiment is selecting the layer functions to determine the detailed structure of the proposed method using two stacked models and reconstructing the 32 × 32 images.

For each layer function, the model is trained on 50 epochs. The method is selected according to the evaluation values. Table 1 shows the experiment results. The evaluation value results at the downscaling process are 0.0107, 0.0099, and 0.0086 for max pooling, linear scaling, and 2 × 2 stride 2 convolution, respectively. Therefore, we adopt the 2 × 2 stride 2 convolution.

Using convolution in the downscaling process, we assume that deconvolution (Zeiler et al., 2011) would apply to the upscaling process, considering the symmetry. However, the results of the respective evaluation values are 0.0086 and 0.0107 for linear scaling and deconvolution, respectively. Therefore, we adopt the linear scaling considering the results of deconvolution.

As a generalization performance improvement method, the proposed method considers enhancing each layer with a dropout one and tests it accordingly. However, we do not adopt it because the results deteriorated.

Hence, we adopt the batch normalization (Ioffe and Szegedy, 2015) because of its effectiveness. The evaluation values are 0.0141 and 0.0083 for dropout and batch normalization, respectively.

The evaluation results are 0.0567, 0.0132, and 0.0086 for the sigmoid function, ReLU, and LekyReLU, respectively, in activation function. The sigmoid function allows achieving acceptable results when the number of stacks is shallow; however, it could not be restored at the shallow stage and does not converge when the stacks are deepened. We observe that ReLU (Nair and Hinton, 2010) and LeakyReLU (Maas et al., 2013) converged. However, LeakyReLU achieves better results, and therefore, we adopt it into the model.

The input and output of U-Net are images. Skip connection is incorporated for each U-Net in each stack because information losses correspond to each stacked U-Net. We also add the previous U-Net input image to the next U-Net image and the first input image to all U-Net input images; however, they do not converge.

Concerning optimization methods, we compare Adam (Kingma and Ba, 2015) and AdaBound (Luo et al., 2019). Convergence results are better in Adam: 0.0142 and 0.0071 for Adabound and Adam, respectively.

We adjust the number of stacks. We also train each model by considering from one to eight U-Net stacks. Table 2 shows the results. The discrepancy between the training and validation losses become severe at epoch 30 for more than five stacks, resulting in overlearning. Therefore, we adopt four stacks in the 32 × 32 images.

We use the dataset from CIFAR-100 (Krizhevsky and Hinton, 2009), a database of various images, for training. By generating a sinogram from the dataset, we could obtain an actual cross-sectional image and a sinogram. The CIFAR-100 contains colored 32 × 32 images; therefore, the images are gray-scaled and clopped in a circle to make
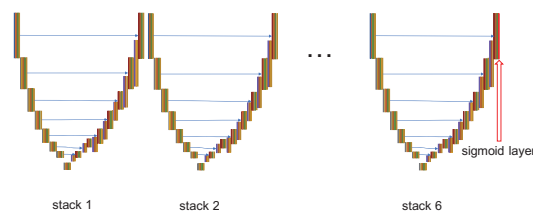

Figure 2: Stacked U-Net

Table 1
Comparison of the applied methods

Class	Method	Adoption	Validation loss
Downscaling	2 × 2 stride 2 convolution	✓	0.0086
	Max pooling		0.0107
	Linear scaling		0.0099
Upscaling	2 × 2 stride 2 deconvolution		0.0107
	Linear scaling	✓	0.0086
Generalization	Batch normalization	✓	0.0083
	Dropout		0.0141
Activate function	Sigmoid		0.0567
	ReLU		0.0132
	LeakyReLU	✓	0.0086
Skip connection	Previous U-Net input image to the next U-Net image		—
	First input image to all U-Net input images		—
Optimizer	Adam	✓	0.0071
	AdaBound		0.0142

Table 2
Relationship between the number of stacks and evaluation values for the 32 × 32 images

Number of stacks	Training loss	Validation loss
1	0.0070	0.0073
2	0.0070	0.0072
3	0.0051	0.0053
4	0.0053	0.0054
5	0.0053	0.0301
6	0.0050	0.0173
7	0.0052	0.0232
8	0.0054	0.0479

the conditions aligned with CT images. We use mean squared error as the evaluation value. Figure 3a–3c shows the resulting image obtained by reconstruction. At that time, the PSNR and SSIM values are 26.82 db and 0.951, respectively.

4.2. Discussion of reconstruction results in 32 × 32 images

Figure 3c shows that the outline and shadow of an apple are represented correctly; however, its stems are not reproduced. The PSNR is also low and insufficient for the reconstruction.

The evaluation values at convergence gradually improve with an increased number of stacks (one to four stacks); however, the evaluation values became worse when that number of stacks is five or more. This can be caused by the lack of information in the input image being transmitted to the later U-Net when the stacks become deeper. We also apply skip connection to convey to the back end; however, there is poor convergence. However, there is a possibility of devising this method, as this reconstruction level is not practical. The amount of information in restoring the original image into the sinogram image itself as input also decreases because of low-resolution reconstruction, limiting the restoration of the detailed structure. Therefore, we shift to 512 × 512 images.

4.3. Reconstruction with 512 × 512 images

Based on the study results in the reconstruction of 32 × 32 images, we shift to studying 512 × 512 images. Here, the model structure varies from three to seven steps, and the number of steps is determined by learning 30 iterations. We adopt six stacked models, resulting in fast convergence.

We perform a comparison experiment with a small number of projections. The constructed neural net only accepts input of 512 × 512; therefore, we apply bicubic processing to upscale the sinogram image to a 512 × 512. Suppose that the number of projections is small; in that case, the sinogram image to be inputted has the size of 64 × 512 (where the number of projections is 64), which could not be inputted. This issue could be solved by applying bicubic processing

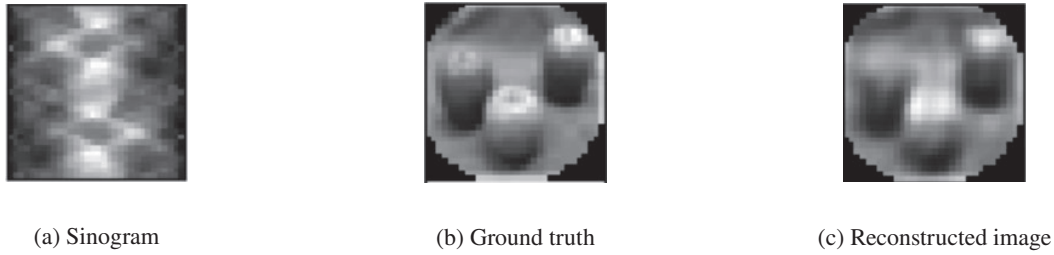| (a) Sinogram | (b) Ground truth | (c) Reconstructed image |

Figure 3: Comparison of the 32 × 32 images

and upscaling.

As high-resolution images are required, we utilize 28,463 images from the image database ImageNet (Krizhevsky et al., 2017) as the training set and 1,537 images as the validation set. We used the Cancer Imaging Archive (Kirk et al., 2016) (Kirk et al., 2016) as the test set, in which we evaluate 467 images.

An adjustment has been performed for consistency with the images from ImageNet because they have different resolutions. Hence, as a preprocessing step, we gray-scale all images, clip the center to a square not to change the aspect ratio, upscale them to 512 × 512 using bicubic processing, and clip them to a circle to obtain the output images (Figure 4b). A sinogram image (Figure 4a) is generated on the basis of the output image. It is used as the input image.

We experiment with a machine running two CPUs (Intel Xeon E5-2680 v4 at 2.4 GHz), 256 GB of RAM, and four GPUs (NVIDIA TESLA P100). The model is trained using 300 iterations for 90 h.

Figure 4b and 4c shows an example of an image with the 512 projections restored by the validation set. Figure 4a depicts the sinogram image that has been inputted to the model, and Figure 4c corresponds to the image outputted by the model. At this time, the PSNR and SSIM values corresponding to the result provided in Figures 4b and 4c are 32.24 db and 0.963, respectively.
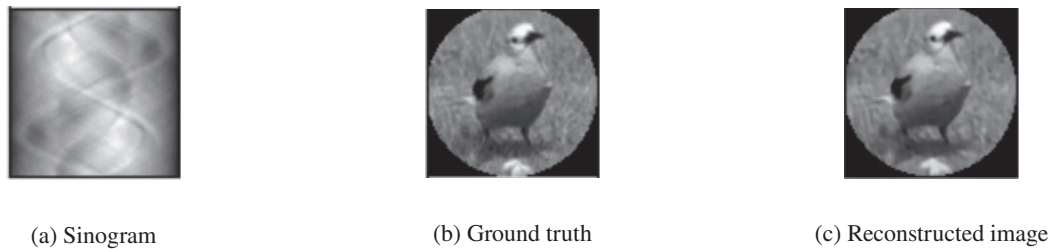

| (a) Sinogram | (b) Ground truth | (c) Reconstructed image |

Figure 4: Results of the validation set

4.4. Results in medical images

Table 3 shows the comparison using the proposed method and the existing methods [FPB, SART, total variation norm (TV-norm), and FBP U-Net in the restoration of medical images. The FBP and SART methods employ the iradon and iradon_sart functions of skimage (der Walt et al., 2014). TV-norm uses the tv1_2d function of proxTV (Alvaro et al., 2018). Jin et al. (2017) showed that FBP U-Net was trained on 101 epochs for 475 medical images with 512 projections (FBP U-Net512) and 64 projections (FBP U-Net64). DD-Net is evaluated using (Zhang, ????).

We experiment with a machine running Intel Core i3-4130T at 2.90 GHz and 24 GB of RAM. We evaluated the performance in terms of PSNR, SSIM, and processing time.

Here, we analyze FBP, SART, and TV-norm methods in terms of the processing speed on the CPU. The proposed method measures the processing time, including the bicubic processing and upscaling. We also estimate the proposed method's processing speed and FBP U-Net and DD-Net on the GPU (GeForce GTX 1080).

There is a constant processing time of the proposed method for all projections: 2.35 s (CPU) and 0.11 s (GPU). The processing times of FBP and SART methods are 0.21 s (CPU) and 2.07 s (GPU). The processing time of FPB-U-Net is 6.54 s (CPU) and 1.02 s (GPU).

The mean PSNR values of the proposed method, FBP, SART, TV-norm, FBP U-Net512, FBP-Net64, and DD-Net for 64 projections are 27.93, 21.41, 25.03, 25.43, 19.79, 22.27, and 22.75 db, respectively. The mean SSIM values of

Table 3
Comparison of the proposed method with the existing methods

Method	Number of projections	PSNR (db)	SSIM	processing time(s) (GPU)
Proposed method	512	29.67	0.907	2.35 (0.11)
	256	28.43	0.894	2.35 (0.11)
	128	28.20	0.891	2.35 (0.11)
	64	27.93	0.886	2.35 (0.11)
FBP	512	35.17	0.961	1.71
	256	31.93	0.901	0.82
	128	26.56	0.734	0.41
	64	21.41	0.547	0.21
SART	512	31.07	0.949	16.79
	256	29.35	0.926	8.30
	128	27.24	0.884	4.11
	64	25.03	0.829	2.07
TV-norm	512	30.56	0.910	2.31
	256	30.50	0.909	1.49
	128	29.86	0.893	1.01
	64	25.73	0.750	0.21
FBP U-Net512	512	26.79	0.858	8.16 (2.70)
	256	25.38	0.798	7.24 (1.72)
	128	22.93	0.687	6.75 (1.26)
	64	19.79	0.581	6.54 (1.04)
FBP U-Net64	512	23.13	0.733	8.06 (2.70)
	256	23.02	0.726	7.14 (1.72)
	128	23.01	0.709	6.76 (1.26)
	64	22.27	0.666	6.54 (1.02)
DD-Net	512	26.75	0.911	6.49 (2.66)
	256	26.58	0.898	5.60 (1.77)
	128	25.51	0.832	5.19 (1.36)
	64	22.75	0.717	4.99 (1.16)

the proposed method, FBP, SART, TV-norm, FBP U-Net512, FBP-Net64, and DD-Net for 64 projections are 0.886, 0.547, 0.829, 0.750, 0.581, 0.666, and 0.717, respectively. The proposed method is superior to the existing methods in 64 projections.

Figure 5a–5h shows the reconstructed results in 64 projections. In Figure 5b, the proposed method does not have any artifacts spreading radially from the center.

Figure 6a–6h shows the upper right-hand details of the cross-sectional images. In the detailed image, the detailed blood vessels in the original image have disappeared in all methods.

The artifacts disappear in the proposed method and FBP U-Net64. The edges of the large structure are well represented by the proposed method, and the thick vessels located in the center are also clearly represented.

In FBP U-Net64, the structure of thick blood vessels is the most common. However, the detailed structure of the large artery's tip near the lower-left corner of the center is not captured in the proposed method. The FPB and TV-norm methods provide more details than the proposed method. Table 4 provides the PSNR and SSIM values for the detailed part.

The PSNR values of the proposed method, FBP, SART, TV-norm, FBP U-Net512 FBP U-Net64 and DD-Net are 30.37, 25.17, 29.34, 29.61, 18.75, 22.92 and 18.03 db, respectively. The SSIM values of the proposed method, FBP, SART, TV-norm, FBP U-Net512, FBP U-Net64, and DD-Net are 0.896, 0.57, 0.863, 0.816, 0.494, 0.447, and 0.510 respectively. Hence, the proposed method is superior to all methods concerning the detailed part.

The detailed part described above corresponds to the vessel area, and the vital part of the CT image of the vessel is used to determine whether the plaque is calcified (appears white) or not. In the proposed method, no large structures are missed or lost in the noise, facilitating distinct classifications.

(a) Ground truth

(b) Proposed method

(c) FBP

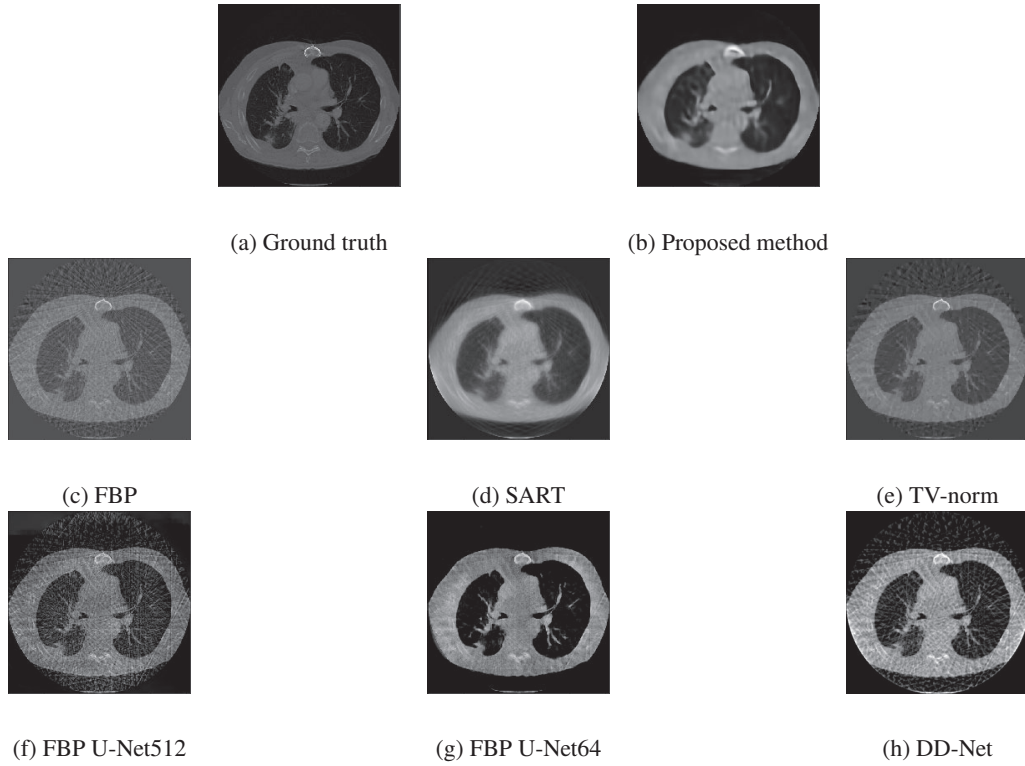(d) SART

(e) TV-norm

(f) FBP U-Net512

(g) FBP U-Net64

(h) DD-Net

Figure 5: Reconstructed images with 64 projections

Table 4
Comparison of details with 64 projections

Method	PSNR (db)	SSIM
Proposed method	30.37	0.896
FBP	25.17	0.570
SART	29.34	0.863
TV-norm	29.61	0.816
FBP U-Net512	18.75	0.494
FBP U-Net64	22.92	0.447
DD-Net	18.03	0.510

4.5. Comparison of the accuracy between the proposed method and FBP U-Net when learning with ImageNet

To compare with the proposed method, we also train the FBP U-Net on the ImageNet data with 64 projections. Figure 7 represents the results of reconstruction on the medical image. The mean PSNR and SSIM values of the restored results are 18.99 db and 0.581, respectively. This result is even lower than that of the medical image training.

Figure 5b shows that the proposed method does not provide any artifacts. Figure 7 shows that the whole area is covered with haze-like artifacts.

4.6. Comparison using segmentation task

It is not easy for medical experts to judge many images, whether they are good or bad. We input the reconstructed images to a model that has already been trained with the segmentation task; then, we compare the results by evaluating whether the generated images can be used for medical purposes or not (Seitzer et al., 2018). We use UNet++ (Zhou et al., 2018) as the model for comparison. We use eighty 512×512 images from MedSeg Covid Dataset 1 (MedSeg;

(a) Ground truth

(b) Proposed method

(c) FBP

(d) SART

(e) TV-norm

(f) FBP U-Net512
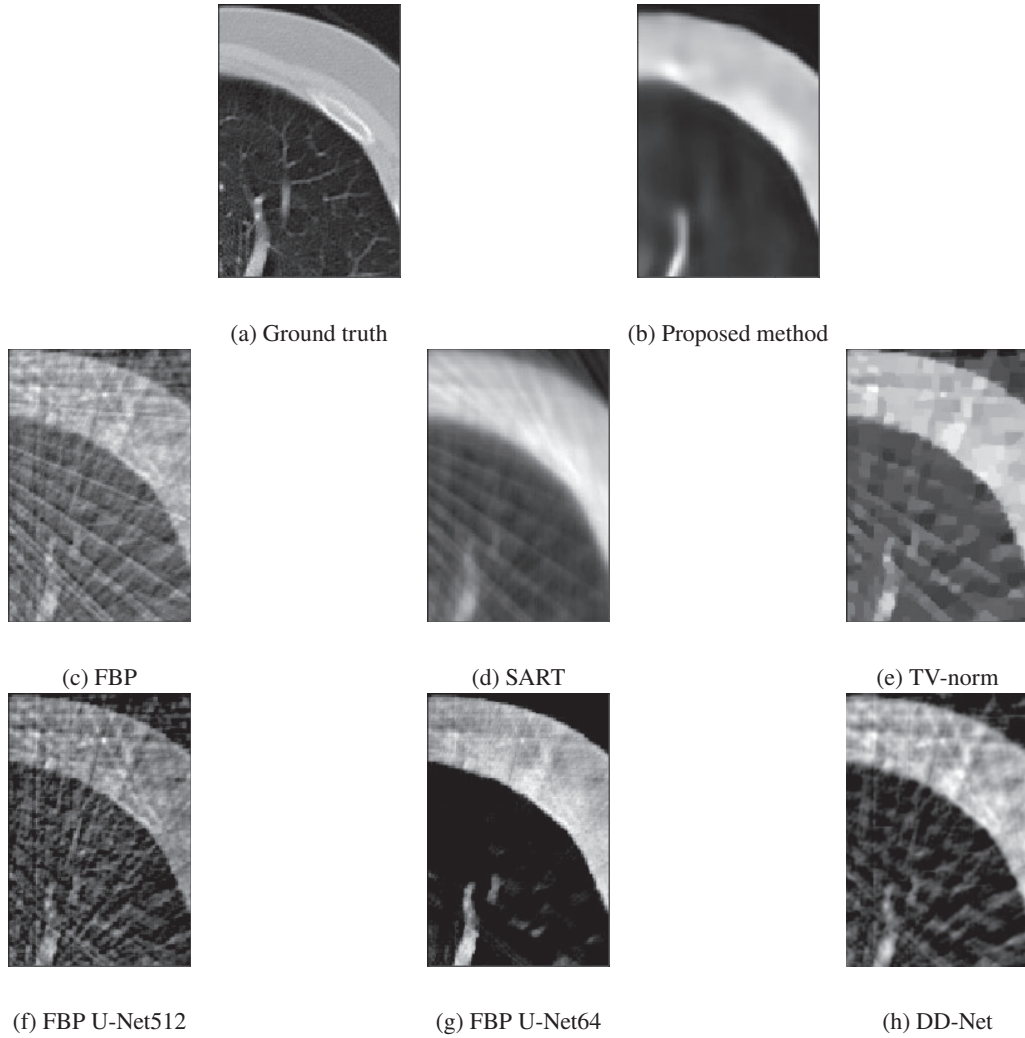
(g) FBP U-Net64

(h) DD-Net

Figure 6: Comparison of details with 64 projections

Figure 7: Results of medical images of FBP U-Net trained on the ImageNet with 64 projections

and Jenssen, Håvard Bjørke; Sakinis, 2021) for training and 20 images for validation; a total of 250 epochs are trained. The trained model has an average Dice score of 0.555 on the validation set. Table 5 provides the average Dice scores for the segmentation task. The proposed method shows the best results with 64 projections, whereas the TV-norm shows the best results for the following projections: 128, 256, and 512.

Table 5
Comparison using segmentation task

Method	Number of projections	Dice
Proposed	512	0.556
	256	0.545
	128	0.540
	64	0.546
FBP	512	0.556
	256	0.552
	128	0.538
	64	0.503
SART	512	0.458
	256	0.462
	128	0.478
	64	0.490
TV-norm	512	0.562
	256	0.562
	128	0.561
	64	0.539
FBP U-Net512	512	0.550
	256	0.541
	128	0.522
	64	0.476
FBP U-Net64	512	0.543
	256	0.544
	128	0.543
	64	0.533
DD-Net	512	0.514
	256	0.502
	128	0.478
	64	0.437

5. Discussion

The results corresponding to the validation set (Figure 5b) demonstrate that even the grass in the background of the considered image of a bird is restored appropriately, indicating that the proposed method performs reconstructions successfully.

However, in the test set, concerning the medical images (Figures 6b and 8), the detailed structure has been lost. This can be explained using various images instead of medical images. Therefore, there is room for improvement of the proposed method's accuracy by adding cross-sectional images to the training dataset.
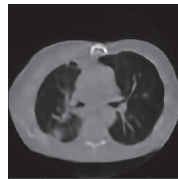

Figure 8: Results of reconstructing medical images with 512 projections

The results of reconstruction by the proposed method are compared with those by FBP U-Net. Figure 5b represents that the proposed method relies on a model trained on various images with 512 projections. However, the amount of noise and artifacts are reduced. Concerning FBP U-Net512, we observe that the noise has not been removed. The restored results of the model trained by FBP U-Net64 using the images of 64 projections are noised (see Figure 5g).

This difference can be explained by the fact that FBP U-Net512 learns using images with 512 projections. Generally, FBP U-Net learns how to eliminate noise according to the number of projections. Therefore, FBP U-Net512

(a) Stack 1 output
(b) Stack 2 output
(c) Stack 3 output
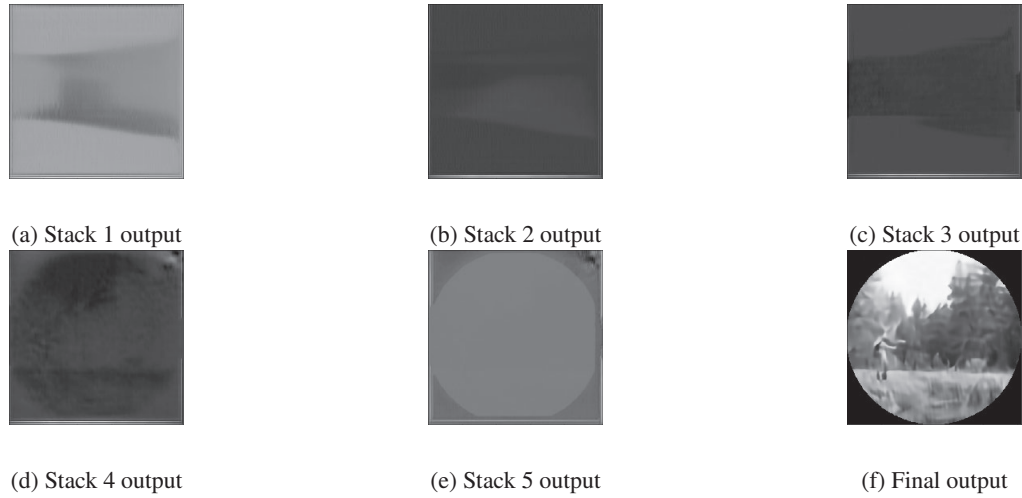(d) Stack 4 output
(e) Stack 5 output
(f) Final output

Figure 9: Intermediate output corresponding to the validation set with 512 projections

cannot remove noise corresponding to 64 unlearned projection images.

The proposed method can perform reconstruction in 64 projections without any noise, although the model has learned with 512 projections. This is because the proposed method has learned how to reconstruct itself. Therefore, the proposed method demonstrates a high generalization performance and does not require relearning the model according to the number of projections.

In contrast, FPB U-Net requires training the model for each number of projections separately. The same model can be used in the proposed method even if the number of X rays projected due to modifying the CT equipment settings has changed. FBP U-Net needs to define a model for each configuration.

The detailed structures are also compared (Figure 6b and 6g). In the figure, the flat area representing the viscera at the top of the image, the proposed method produces a smooth image with no artifacts. In contrast, for FBP U-Net64, there are linear artifacts that cannot be removed. In the area at the lower left of the image representing large vessels, the images are clearly visible in both the proposed method and FBP U-Net64. Concerning large vessels, we have found that FBP U-Net64 can reconstruct a more detailed structure than the proposed method. Comparing the fine arteries in the bottom right of the image shows that the vessels have disappeared in the proposed method. In contrast, in the result corresponding to FBP U-Net, approximately four large noises are visible. The aforementioned difference is expressed as a discrepancy in PSNR and SSIM between the considered methods.

Figure 9a–9f represents the intermediate output results for each stage for the case with 512 projections in the validation set. Figure 9a–9c represents the figures on the cone stretched horizontally. Figure 9d shows the shape of the circle. Figure 9e depicts a clearer shape of the circle. Figure 9f provides the final output image.

In general iterative reconstruction, the shape of a cross-sectional image gradually becomes more apparent, as shown in Figure 10.

Figure 10: Reconstruction of a phantom by the SART method

The intermediate image in the proposed method is reconstructed differently compared with the iterative reconstruction method. The proposed method performs reconstruction in a mode that differs from the conventional model.

In terms of the execution speed using CPU, the proposed method is superior to the SART method in processing

time with up to 128 projections; even at 64 projections, the processing times of the proposed method and SART are 2.35 and 2.07 s, respectively.

In GPU, the proposed method outperforms the existing methods in all cases. As GPUs' hardware architecture is generally different from that of CPUs, it is necessary to change the algorithms and implementations accordingly to run processing appropriately (Vlček, 2005). As deep learning combines linear computational processing steps, the proposed method can run the same GPU model.

The proposed method has no artifacts spreading radially from the center and no breakage in image quality, achieving a high generalization performance with 64 projections. Although we have not utilized images with a reduced number of projections for training, the proposed method can recover unknown inputs without failures.

Concerning different numbers of projections that have not been considered for learning, we have found that the proposed method reduces PSNR by approximately 0.30 db even if the number of projections is reduced by half, indicating a high generalization performance for the unknown data. When the number of projections is reduced by half in the SART method, a decrease in PSNR is equal to 2.0 db, indicating that the proposed method achieves better results when the number of projections is lower.

Concerning the detailed part of the reconstructed image, we have found that the proposed method outperforms all considered methods. This can be explained by the fact that the proposed method does not produce any artifacts, confirming the proposed method's effectiveness in diagnosis, including determining calcification that requires more detailed representation in the vascular area.

However, when analyzing the details of an image obtained by the proposed method, the overall image is flat, and the high-frequency components are not captured.

In contrast, in the FBP method, the detailed structure can be seen in the noise. This is because the proposed method does not consider medical images for learning and, therefore, does not fully capture the features required to represent medical images. A performance improvement can be achieved by mixing several medical images with those in the training dataset.

The proposed method shows the best results with 64 projections in the segmentation task, outperforming the deep learning methods (FBP UNet-64, FBP UNet-512, and DD-Net) in almost all projections (128, 256, and 512). The best results are obtained when the TV-norm is between 128 and 512 projections, which can be attributed to the sparsity of the TV-norm in segmentation, preserving the information of essential regions. The SART method is resistant to noise and artifacts; however, it produces not good images in segmentation. Hence, the proposed method is superior because it simultaneously removes noise and artifacts and retains region information. The Dice score of the segmentation task model is 0.555 because of the lack of learning. Therefore, it is necessary to create a segmentation task model with higher accuracy in the future.

In deep learning, the larger a training dataset is, the more accurate is the obtained result (Cho et al., 2015; Figueroa et al., 2012). Therefore, we expect that if the proposed method is trained using images with a low number of projections, the reconstruction accuracy for a smaller number of projections will be improved.

The method itself also has room for improvement in terms of accuracy. Guo et al. (2020) applied Attention block (Bahdanau et al., 2015) to MRI reconstruction with GAN. When applied to X-ray CT image reconstruction, Attention focuses on a specific range of pixels and determines their importance. This is incompatible with transforming a sinogram image into a cross-sectional image, which requires looking at the entire image to solve. When combined with the recurrent model, Attention can retain the global context; therefore, the essential information for the recovered pixels can be selected from the entire image, improving the accuracy. Schlemper et al. (2018) applied DC-CNN (Schlemper et al., 2018) and stochastic depth (Huang et al., 2016) for MRI reconstruction. They achieved better results than those existing methods. Stochastic depth cannot be applied to the proposed method because it requires ReLU with positive output values, which employs LekyReLU. In this experiment, the training time is long, and the results deteriorate when the number of stacks increases. By applying connection control like stochastic depth, we expect to reduce the training time and improve accuracy by increasing the number of stacks. Zhu et al. (2019) combined the lesion's ROI and GAN to create an accurate model with fast learning convergence. It is not possible to identify the lesion because the proposed method recovers directly from the sinogram. However, there is a possibility that the accuracy can be improved by identifying lesions and creating ROIs after applying our method and then applying GAN. If the lesion's ROI can be recovered directly from the sinogram, the computational cost of the subsequent steps can also be reduced.

6. Conclusion

Based on the stacked U-Net model, the proposed method can reconstruct a cross-sectional image by adjusting the layer function and the number of stacks. This method can also reconstruct a projected image requiring a small number of projections and image quality comparable with the iterative reconstruction method. Moreover, the proposed method is advantageous in terms of the high processing speed. The proposed method can be directly utilized to reconstruct the cross-sectional image without medical images. This feature allows avoiding the problem of reproducing the noise associated with the FBP method, which is a problem of the existing FPB-based U-Net method.

In the proposed method, the processing time for reconstruction is 2.35 s, equivalent to that of the existing method. However, it is also easier to run on GPU than the existing method (0.11 s). As reconstruction is possible even on a small number of projections, the input image with a small number of projections can output a cross-sectional image with fewer artifacts than the existing methods. By reducing the number of projections, we have found that the volume of X rays can be limited accordingly, and the patients' burden can be decreased.

However, applying the current model results in a loss of the structure in a detailed view. Therefore, improvement of the accuracy is a problem to be considered in future-related research. This issue may be observed because of the lack of appropriate data for training and relevant medical image knowledge incorporated into the model. We will improve the accuracy by extending the training dataset by including the images with fewer projections, increasing the number of images considered for training, and mixing medical images with the training images. We will also examine the effects of the recently proposed structures to improve accuracies, such as Attention and ResBlock, and methods such as stochastic depth, and improve the model itself.

In this study, the evaluation of images has been performed quantitatively. For applying the proposed method to the actual medical practice, it is necessary to make the images easy for doctors to diagnose. In the future, we will ask physicians to evaluate the results and specify the extent to which they can be used in the medical field.

In this study, the results have deteriorated when the number of stacks has been deepened. This may be because an input image's features are not transmitted to the output appropriately when the stacks are deep. We consider that it is necessary to propose a structure in which such an aggravation of the results does not occur. Here, we have demonstrated that the proposed method can be applied to reconstruct differently using the SART methods. By clarifying how this method can be used for reconstruction, we will improve the proposed method's accuracy and existing methods.

7. Acknowledgments

References

A. C. Kak, M. Slaney, Principles of Computerized Tomographic Imaging, IEEE Press, 2001. doi:10.1137/1.9780898719277.

A. Andersen, Ultrasonic Imaging 6 (1984) 81–94. doi:10.1016/0161-7346(84)90008-7.

A. H. Andersen, IEEE Transactions on Medical Imaging 8 (1989) 50–55. URL: http://www.ncbi.nlm.nih.gov/pubmed/18230499. doi:10.1109/42.20361.

A. P. Dempster, N. M. Laird, D. B. Rubin, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1977) 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x.

L. L. Geyer, U. J. Schoepf, F. G. Meinel, J. W. Nance, G. Bastarrika, J. A. Leipsic, N. S. Paul, M. Rengo, A. Laghi, C. N. De Cecco, Radiology 276 (2015) 339–357. URL: www.rsna.org/rsnarights. doi:10.1148/radiol.2015132766.

H. M. Hudson, R. S. Larkin, IEEE Transactions on Medical Imaging 13 (1994) 601–609. doi:10.1109/42.363108.

D. Kim, S. Ramani, J. A. Fessler, IEEE Transactions on Medical Imaging 34 (2015) 167–178. doi:10.1109/TMI.2014.2350962.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Neural Computation 1 (1989) 541–551. doi:10.1162/neco.1989.1.4.541.

A. Krizhevsky, I. Sutskever, G. E. Hinton, Communications of the ACM 60 (2017) 84–90. doi:10.1145/3065386.

S. Lefkimmiatis, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, pp. 5882–5891. doi:10.1109/CVPR.2017.623.

W. S. Lai, J. B. Huang, N. Ahuja, M. H. Yang, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 5835–5843. doi:10.1109/CVPR.2017.618. arXiv:1704.03915.

O. Ronneberger, P. Fischer, T. Brox, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 9351, Springer Verlag, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28. arXiv:1505.04597.

Y. Yang, J. Sun, H. Li, Z. Xu, in: Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 10–18. URL: https://papers.nips.cc/paper/6406-deep-admm-net-for-compressive-sensing-mri.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Foundations and Trends in Machine Learning 3 (2010) 1–122. doi:10.1561/2200000016.

G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, D. Firmin, IEEE Transactions on Medical Imaging 37 (2018) 1310–1321. doi:10.1109/TMI.2017.2785879.

M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, 2014. URL: http://arxiv.org/abs/1411.1784. arXiv:1411.1784.

J. Zhu, G. Yang, P. Lio, in: E. D. Angelini, B. A. Landman (Eds.), Medical Imaging 2019: Image Processing, volume 10949, SPIE, 2019, p. 56. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10949/2512576/Lesion-focused-super-resolution/10.1117/12.2512576.full. doi:10.1117/12.2512576.

P. Zhu, Jin; Guang, Yang; Pedro, Ferreira; Andrew, Scott, Sonia, Nielles-Vallespin; Jennifer, Keegan; Dudley, F. Pietro, Lio; David, in: In the International Society for Magnetic Resonance in Medicine 27th Annual Meeting, p. 1. URL: https://archive.ismrm.org/2019/0778.html.

J. Zhu, G. Yang, P. Lio, in: Proceedings - International Symposium on Biomedical Imaging, volume 2019-April, IEEE Computer Society, 2019, pp. 1669–1673. URL: http://arxiv.org/abs/1901.03419. arXiv:1901.03419.

S. Yu, H. Dong, G. Yang, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, D. Firmin, Y. Guo, arXiv (2017). URL: http://arxiv.org/abs/1705.07137. arXiv:1705.07137.

K. H. Jin, M. T. McCann, E. Froustey, M. Unser, IEEE Transactions on Image Processing 26 (2017) 4509–4522. doi:10.1109/TIP.2017.2713099. arXiv:1611.03679.

J. Kim, J. K. Lee, K. M. Lee, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, IEEE Computer Society, 2016, pp. 1646–1654. doi:10.1109/CVPR.2016.182.

K. He, X. Zhang, S. Ren, J. Sun, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, Microsoft, pp. 770–778. URL: http://image-net.org/challenges/LSVRC/2015/. doi:10.1109/CVPR.2016.90. arXiv:1512.03385.

Z. Zhang, X. Liang, X. Dong, Y. Xie, G. Cao, IEEE Transactions on Medical Imaging 37 (2018) 1407–1417. doi:10.1109/TMI.2018.2823338.

A. Sevastopolsky, S. Drapak, K. Kiselev, B. M. Snyder, J. D. Keenan, A. Georgievskaya, Medical Imaging 2019: Image Processing 10949 (2019) 78. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10949/2511572/Stack-U-Net--refinement-network-for-improved-optic-disc/10.1117/12.2511572.full. doi:10.1117/12.2511572. arXiv:arxiv.org/abs/1804.11294.

S. Shah, P. Ghosh, L. S. Davis, T. Goldstein (2018). URL: http://arxiv.org/abs/1804.10343. arXiv:1804.10343.

S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, volume 2017-July, IEEE Computer Society, 2017, pp. 1175–1183. doi:10.1109/CVPRW.2017.156. arXiv:1611.09326.

A. Newell, K. Yang, J. Deng, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 9912 LNCS, Springer Verlag, 2016, pp. 483–499. doi:10.1007/978-3-319-46484-8_29. arXiv:1603.06937.

Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.

M. D. Zeiler, G. W. Taylor, R. Fergus, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2018–2025. doi:10.1109/ICCV.2011.6126474.

S. Ioffe, C. Szegedy, in: 32nd International Conference on Machine Learning, ICML 2015, volume 1, pp. 448–456. arXiv:1502.03167.

V. Nair, G. E. Hinton, in: ICML 2010 - Proceedings, 27th International Conference on Machine Learning, pp. 807–814.

A. L. Maas, A. Y. Hannun, A. Y. Ng, in: ICML Workshop on Deep Learning for Audio, Speech and Language Processing.

D. P. Kingma, J. L. Ba, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. arXiv:1412.6980.

L. Luo, Y. Xiong, Y. Liu, X. Sun, in: 7th International Conference on Learning Representations, ICLR 2019. URL: https://github.com/Luolc/AdaBound. arXiv:1902.09843.

A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images., 2009.

S. Kirk, Y. Lee, C. Roche, E. Bonaccio, J. Filippini, R. Jarosz, Radiology Data from The Cancer Genome Atlas Thyroid Cancer [TCGA-THCA] collection, 2016. doi:10.7937/K9/TCIA.2016.9ZFRVF1B.

S. der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, PeerJ 2 (2014) e453.

B. Alvaro, Suvrit, Sra, Journal of Machine Learning Research 19 (2018) 1–82. URL: http://jmlr.org/papers/v19/13-538.html.

Z. Zhang, zzc623/DD_Net: code for the paper entitled "A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution", ???? URL: https://github.com/zzc623/DD{_}Net.

M. Seitzer, G. Yang, J. Schlemper, O. Oktay, T. Würfl, V. Christlein, T. Wong, R. Mohiaddin, D. Firmin, J. Keegan, D. Rueckert, A. Maier, in: A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, 2018, pp. 232–240.

Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A Nested U-Net Architecture for Medical Image Segmentation: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings, volume 11045, 2018, pp. 3–11. doi:10.1007/978-3-030-00889-5_1.

MedSeg;, T. Jenssen, Håvard Bjørke; Sakinis, MedSeg Covid Dataset 1, 2021. doi:https://doi.org/10.6084/m9.figshare.13521488.v2.

V. Vlček, volume 4, WSEAS Korfu, 2005, pp. 34–39.

J. Cho, K. Lee, E. Shin, G. Choy, S. Do, in: undefined, ICLR 2016, 2015. URL: http://arxiv.org/abs/1511.06348. arXiv:1511.06348.

R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, L. H. Ngo, BMC Medical Informatics and Decision Making 12 (2012) 8. URL: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-8. doi:10.1186/1472-6947-12-8.

Y. Guo, C. Wang, H. Zhang, G. Yang, arXiv (2020) 2006.12915. URL: http://arxiv.org/abs/2006.12915. arXiv:2006.12915.

D. Bahdanau, K. H. Cho, Y. Bengio, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2015. URL: https://arxiv.org/abs/1409.0473v7. arXiv:1409.0473.

J. Schlemper, G. Yang, P. Ferreira, A. Scott, L. A. McGill, Z. Khalique, M. Gorodezky, M. Roehl, J. Keegan, D. Pennell, D. Firmin, D. Rueckert, in: The 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018), volume 11070 LNCS, Springer Verlag, 2018, pp. 295–303. URL: https://doi.org/10.1007/978-3-030-00928-1{_}34. doi:10.1007/978-3-030-00928-1_34. arXiv:1805.12064.

G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Weinberger, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9908 LNCS (2016) 646–661. URL: http://arxiv.org/abs/1603.09382. arXiv:1603.09382.