Deep Learning–Based Image Noise Quantification Framework for Computed Tomography

Nathan R. Huber, PhD, Jiwoo Kim,† Shuai Leng, PhD,* Cynthia H. McCollough, PhD,* and Lifeng Yu, PhD**

Objective: Noise quantification is fundamental to computed tomography (CT) image quality assessment and protocol optimization. This study proposes a deep learning–based framework, Single-scan Image Local Variance EstimatoR (SILVER), for estimating the local noise level within each region of a CT image. The local noise level will be referred to as a pixel-wise noise map.

Methods: The SILVER architecture resembled a U-Net convolutional neural network with mean-square-error loss. To generate training data, 100 replicate scans were acquired of 3 anthropomorphic phantoms (chest, head, and pelvis) using a sequential scan mode; 120,000 phantom images were allocated into training, validation, and testing data sets. Pixel-wise noise maps were calculated for the phantom data by taking the per-pixel SD from the 100 replicate scans. For training, the convolutional neural network inputs consisted of phantom CT image patches, and the training targets consisted of the corresponding calculated pixel-wise noise maps. Following training, SILVER noise maps were evaluated using phantom and patient images. For evaluation on patient images, SILVER noise maps were compared with manual noise measurements at the heart, aorta, liver, spleen, and fat.

Results: When tested on phantom images, the SILVER noise map prediction closely matched the calculated noise map target (root mean square error <8 Hounsfield units). Within 10 patient examinations, SILVER noise map had an average percent error of 5% relative to manual region-of-interest measurements.

Conclusion: The SILVER framework enabled accurate pixel-wise noise level estimation directly from patient images. This method is widely accessible because it operates in the image domain and requires only phantom data for training.

Key Words: computed tomography, deep learning, image quality, noise quantification

(*J Comput Assist Tomogr* 2023;47: 603–607)

Computed tomography (CT) is a medical imaging modality that uses x-ray radiation to obtain a 3-dimensional representation of human anatomy. Computed tomography image quality assessment is performed routinely for equipment evaluation and scanning protocol optimization. One important indicator of image quality is image noise. Noise is typically measured using standardized image quality phantoms; however, phantom-based measurement is not ideal because it does not reflect how the system operates on patients in standard practice. Noise measurement techniques within patient examinations are limited; most commonly, noise in patient examinations is manually measured as the SD of CT numbers within a uniform region of interest (ROI SD). Ideally, there would be fully automatic tools for measuring pixel-wise noise level in patient CT images. The difficulty of reliable noise quantification in patient images is a barrier for protocol optimization and image quality standardization across patients and practices.

Some methods have been proposed for global and pixel-wise measurement of image noise in patient CT images. Global metrics aim to distill the noise level within a patient examination into a single quantity.[1–7] Christianson et al[1] described a global noise index, which automatically determines uniform regions of patient anatomy, applies SD measurements in these regions, and reports the most frequent noise level measured. Global noise assessment has also been achieved with deep learning–based methods; a convolutional neural network (CNN) was trained to predict radiologist-assigned labels of subjective image quality ratings for patient CT images[8,9] and a generative adversarial network as trained to predict patient-specific noise power spectrum.[10] In contrast, pixel-wise noise quantification aims to quantify the spatial variations in noise level within individual patient images. For simple geometries, pixel-wise noise characteristics can be analytically determined by propagating a noise model through the reconstruction process.[11,12] To mimic a clinical scenario, CT simulation tools and projection noise insertion can be used to approximate pixel-wise CT patient noise.[13] However, previous techniques for pixel-wise noise quantification have not been adopted because of inaccessibility of clinical CT projection data, lack of manufacturer transparency about the data preprocessing and image reconstruction process, extensive computational processing times, or inaccuracy of the results.

In this study, we propose a deep learning–based method to estimate the pixel-wise noise level of patient CT images; we refer to this technique as Single-scan Image Local Variance EstimatoR (SILVER). Based on prior work demonstrating CNN for noise reduction,[14,15] we hypothesized that a CNN would be capable of predicting pixel-wise noise level.

MATERIALS AND METHODS

Training Data Set

SILVER was trained using CT images of 3 different anthropomorphic phantoms, which mimicked the body habitus of the head (Angiographic CT Head Phantom ACS; Kyoto Kagaku, Kyoto, Japan), chest (LUNGMAN; Kyoto Kagaku), and pelvis (RSD Sectional Phantom; Maplewood, MN). One hundred replicate scans of each phantom were performed in a sequential scan mode using a dual-source 128-slice scanner (Somatom Definition Flash; Siemens Healthineers, Erlangen, Germany) at 120 kV with routine dose (RD; 200 effective mAs) and quarter dose (QD; 50 effective mAs). Automatic tube current and potential systems were turned off for this study. Reconstruction was performed with a smooth

From the *Department of Radiology, Mayo Clinic, Rochester, MN; and †Columbia University, New York, NY.

Received for publication September 14, 2022; accepted February 2, 2023.

Correspondence to: Lifeng Yu, PhD, 200 First St SW, Rochester, MN 55905 (e-mail: yu.lifeng@mayo.edu).

This work was supported by the CT Clinical Innovation Center, the Mayo Clinic Graduate School of Biomedical Sciences, the Mayo Clinic Summer Undergraduate Research Fellowship, and the National Institutes of Health under award numbers U01 EB017185 and U24 EB028936. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research support is provided to the Mayo Clinic from Siemens AG, unrelated to this work.

The authors declare no conflict of interest.

DOI: 10.1097/RCT.0000000000001469

kernel (Siemens B30, MTF 10% of 5.9, no edge enhancement) and a medium sharp kernel (Siemens D45, MTF 10% of 9.4, contains edge enhancement), an image thickness of 1 mm, and a field of view of 420 mm to match the reconstruction parameters of the American Association of Physicists in Medicine and Mayo Clinic Grand Challenge patient data set, which were used for evaluation (Patient Image Data section).[16] A total of 120,000 CT images of the phantoms were acquired, which were allocated into data sets for model training, validation, and testing (80% for training, 10% for validation, and 10% for testing).

One hundred thousand training patches (64 × 64 pixels) were extracted from phantom images to be used as input (phantom CT image with noise scaling) and target (corresponding pixel noise map). To improve diversity in the training data set, a random linear scaling of image noise (ranging from 0% to 200%) was applied to each patch. Linear noise scaling was applied by subtracting an individual phantom CT image by the 100-repetition average, multiplying the noise-only difference image by a random scaling factor, and then adding the scaled noise-only difference image back into the 100-repetition average (Equation 1). The pixel noise map label was calculated as the pixel-wise SD of each set of 100 repeated phantom images while accounting for the linear noise scaling term (Equation 2),

Training input with noise scaling:

$$f(x_{i,j},\alpha) = x_{i,j} + \alpha(x_{i,j} - x_{i,j}) \quad (1)$$

Training target:

$$\mathrm{SD}[f(x_{i,j},\alpha)] = \sqrt{\frac{\sum(f(x_{i,j},\alpha) - x_{i,j})^2}{n-1}} = \alpha\sqrt{\frac{\sum(x_{i,j} - x_{i,j})^2}{n-1}} \quad (2)$$

where $x_{i,j}$ is the CT number of the pixel at location of (i,j) of the image, $x_{i,j}$ is the pixel average from repeated phantom scans, α is a random noise scaling factor (0%–200%), and n is the number of repeat phantom scans (100).

Training Procedure

A CNN was trained via supervised learning to map phantom CT images (with noise scaling) to a corresponding calculated pixel noise map. The CNN resembled a U-Net architecture.[17] Encoding units consisted of 2-dimensional convolutional layers, batch normalization, rectified linear unit activation, and max pooling. Decoding units consisted of 2-dimensional convolutional layers, batch normalization, rectified linear unit activation, and up-sampling (Fig. 1). Mean-square-error loss function was used with respect to the calculated noise map. During training, rotational data augmentation was applied. SILVER was trained twice, once for smooth kernel (B30) and once for medium-sharp kernel (D45). This is to ensure optimal performance for each kernel. When applying the model to a certain image, we make sure that the correct weights are selected for that reconstruction kernel. In addition, training data from both dose levels were put together to train a single set of weights. The phantom data used for training were acquired at the same mAs setting as the patient examinations used for testing; however, because of variations in patient size there were naturally more variations in noise level of the patient data relative to the phantom data. Therefore, we applied a noise scaling technique to the phantom data to augment our training set to encompass a large range of noise levels. By using a large range of noise levels in the training, we expect it to perform well when applied to the various patient data sets used during testing. Training was conducted using an Nvidia GTX 1080 GPU (Nvidia, Santa Clara, CA) equipped with TensorFlow (Mountain View, CA) and Keras (Mountain View, CA).

Performance Evaluation

Anthropomorphic Phantom Image Data

SILVER was first evaluated using anthropomorphic phantom images that were excluded from the training process. The phantom test data set was also acquired with 100 repetitions so that pixel noise map could be calculated, as described in the Training Data Set section. SILVER was applied to full phantom CT images (512 × 512 pixels), and the predicted noise map was compared directly to the calculated noise map. Root mean square error (RMSE), difference images, and percent error maps of the predicted noise map relative to the calculated noise map were

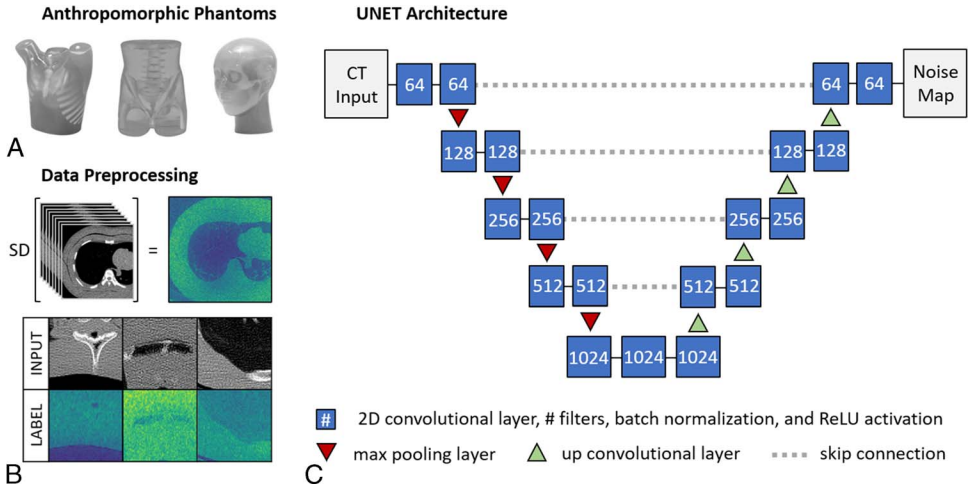

FIGURE 1. Schematic of SILVER framework. A, Chest, pelvis, and head anthropomorphic phantoms were scanned with 100 repeated acquisitions. B, Calculated noise maps were generated as the pixel-wise SD of phantom images. Phantom scans (input) and calculated noise maps (labels) were split into training patches. C, A CNN resembling U-Net was trained to predict a pixel noise map directly from a single CT image. 2D indicates 2-dimensional. This figure can be viewed in color online at www.jcat.org.

used to assess performance. Absolute percent error was defined as the difference between the predicted and calculated noise map, divided by calculated noise map, and multiplied by 100% (Equation 3).

Percent error map$_{i,j}$

$$= \left| \frac{\text{SILVER}[x_{i,j}] - \sqrt{\frac{\sum (x_{i,j} - x_{i,j})^2}{n-1}}}{\sqrt{\frac{\sum (x_{i,j} - x_{i,j})^2}{n-1}}} \right| \times 100\% \qquad (3)$$

where $x_{i,j}$ is the CT number of the pixel at location of (i,j) of the image, $x_{i,j}$ is the pixel average from repeated phantom scans, n is the number of repeated scans (100), and SILVER[$x_{i,j}$] is the predicted noise level at (i,j) from a single scan.

Patient Image Data

SILVER was used to predict pixel-wise noise maps in 10 patient CT data sets from the American Association of Physicists in Medicine and Mayo Clinic Low Dose Grand Challenge data set.[18] This data set contains patient examinations at RD and quarter dose (QD). Quarter dose patient examinations were synthesized using a validated projection-based noise insertion technique, which considers the effect of automatic exposure control, bow tie filter, and electronic noise.[19–21] Manual ROI SD measurements were performed at the aorta, liver, spleen, fat, and heart and compared directly with noise levels predicted by SILVER at the same locations. The ROI radius was set to 10 pixels (8 mm). Twenty-five uniform regions were preselected within each of the 10 data sets. Regions of interest were placed by a physics researcher with 1-year experience. The absolute percent error of each measurement was recorded between SILVER and the ROI SD measurement, and average absolute percent error was calculated for each anatomy.

RESULTS

Anthropomorphic Phantom Image Data

When applied to the anthropomorphic phantom test data set, the noise map predicted by SILVER closely matched the calculated noise map. The RMSE of SILVER noise map relative to the calculated noise map for the test set of each phantom is included in Table 1. For smooth kernel (B30), the average RMSE of the noise map prediction was 1.1 Hounsfield units (HU) at RD and 1.7 HU at QD. For medium sharp kernel (D45), the average RMSE of the noise map prediction was 2.4 HU for RD and 4.7 HU for QD. In general, the SILVER noise map was most accurate within largely uniform regions and less accurate for detailed structures (ie, phantom lung structure). We observed increased error in regions containing streak artifact (ie, phantom heart and chest wall). SILVER performed well for both QD and RD examinations in terms of RMSE, visual inspection, and percent error calculation (Fig. 2).

TABLE 1. Root Mean Square Error (in HU) Between the SILVER Noise Map Prediction and the Calculated Noise Map for the Test Data Set of 3 Anthropomorphic Phantoms (Head, Chest, and Pelvis)

	Smooth Kernel (B30)		Medium Sharp Kernel (D45)	
Phantom	**RD**	**QD**	**RD**	**QD**
Head	0.7	1.1	1.6	2.9
Chest	1.1	1.4	2.1	3.8
Pelvis	1.5	2.5	3.5	7.5
Average	1.1	1.7	2.4	4.7

RMSE was calculated for each phantom at RD and QD.

Patient Image Data

SILVER was used to predict noise maps of 10 patient examinations for 2 dose levels (RD and QD) and 2 reconstruction kernels (B30 and D45). By visual inspection (Fig. 3), SILVER noise prediction matched trends expected regarding patient size (elevated noise observed in large patients), tissue type (elevated noise in bone relative to soft tissue), and depth of region (elevated noise in centermost regions). The accuracy of SILVER noise map was confirmed by comparing to uniform ROI SD measurements (10-pixel radius) at the aorta, liver, spleen, fat, and heart. The absolute percent error of SILVER relative ROI SD measurement is provided for each anatomy in Table 2.

DISCUSSION

In this article, we introduce the SILVER for pixel-wise noise quantification in CT images. The technique was evaluated in 2 ways: (*a*) accuracy of SILVER noise map compared with calculated noise map of anthropomorphic phantom scans and (*b*) accuracy of SILVER noise map compared with ROI SD in 10 patient image cases.

SILVER noise map prediction closely matched the calculated noise map in anthropomorphic phantoms (<8 HU RMSE). SILVER had elevated error within fine phantom lung structures (roughly 50% overestimate). In regions of extensive streak artifact, SILVER tended to underestimate the noise level (roughly 30% underestimate); this may be attributable to insufficient training examples of streak artifacts within our anthropomorphic phantom data set. In the future, we plan to include more phantom geometries within the training data set to improve robustness of SILVER to patient cases containing streak artifact.

SILVER noise map prediction closely matched ROI SD measurements (10-pixel radius) in 10 patient CT data sets. Large uniform structures, such as liver and spleen, achieved the lowest percent error (5% for RD, 4% for QD). The aorta had slightly elevated percent error (7% for RD, 6% for QD). Elevated error at the aorta may be due to streak artifact from the vertebrae. In some patients, the aorta was only slightly larger than the 10-pixel radius; it is possible that nonuniformity in ROI SD measurement impacted the accuracy of these measurements.

There are several notable contributions of this study. To the best of our knowledge, this is the first study using deep learning to quantify pixel-wise noise level directly from CT images. SILVER achieved high accuracy and superior measurement repeatability relative to ROI SD measurement. In addition, we used a widely accessible phantom-based training methodology. Because this technique operates within the image domain, this framework can be implemented on any CT scanner. Because our algorithm is trained directly on CT phantom measurements, it can learn, in theory, the many complexities of CT noise (Poisson and electronic noise sources, internal data processing, geometric variations, and the reconstruction process).

There are several limitations to this study. First, our primary evaluation within patient examinations was based on ROI SD measurements at uniform anatomy. Unfortunately, there are no reliable methods for calculating the noise level within nonuniform patient regions; hence, we could not assess the accuracy of SILVER at these regions. In future work, we plan to perform a cadaver study to assess the accuracy of SILVER noise predictions

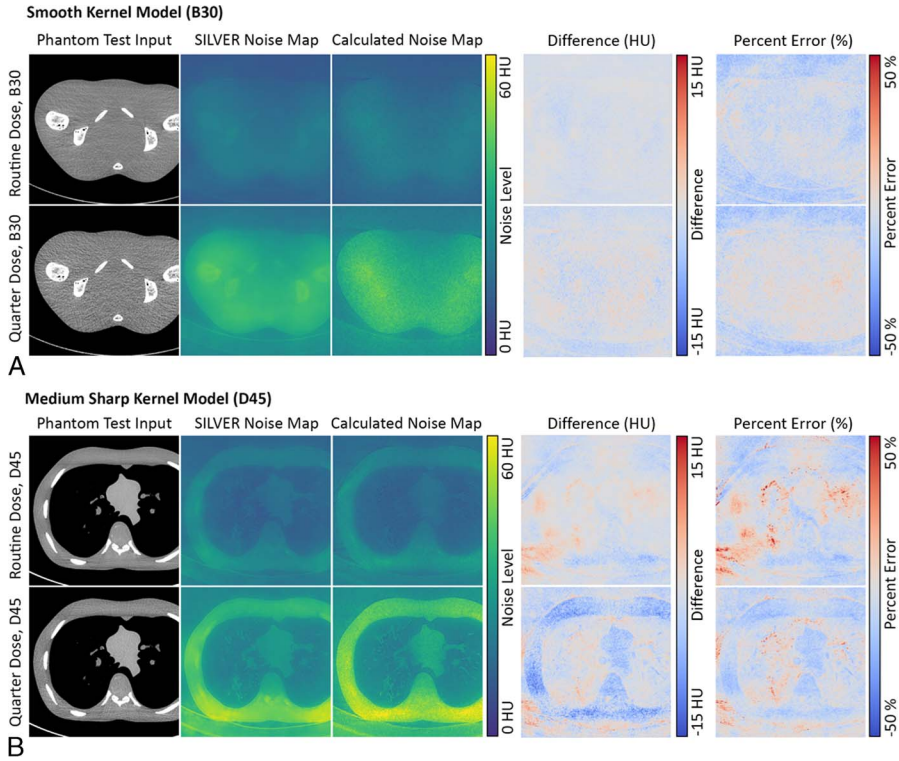

FIGURE 2. SILVER predicted noise map for pelvis and chest phantom for (A) smooth (B30) and (B) medium-sharp kernel (D45) images. The first column is the phantom test CT image (window level: 50, window width: 400); the second column is SILVER noise map prediction; the third column is the calculated noise map based on 100 repeated phantom scans; the fourth column is the difference of SILVER prediction and calculated noise map; and the fifth column is the percent error of SILVER prediction relative to calculated noise map. Image selection was made to show performance within lungs and pelvis. This figure can be viewed in color online at www.jcat.org.

for nonuninform regions of human anatomy. Second, the ability of SILVER to generalize to human anatomic features is dependent on our ability to include similar features within the anthropomorphic phantom training data. For example, we observed reduced accuracy at regions with streak artifact, likely due to insufficient representation of this artifact within the phantoms used for training. In future work, we will explore additional data augmentation techniques to improve upon model generalizability. Third, we demonstrated only SILVER noise maps for filtered back-projection images and a limited number of reconstruction conditions. With retraining, the framework described should also be applicable to iterative reconstruction and other modifications to reconstruction conditions. Further validation is underway. Fourth, this study focused only on prediction of first-order noise level. In future work, we plan to quantify noise correlation directly from patient CT images using a related framework.

The current study was trained and tested on a reconstruction kernel-specific basis (kernel of the training and testing images were matched). Prior literature indicates that CNN models perform poorly when applied to images reconstructed differently than the training data set due to differences in the spatial frequencies of noise.[22,23] We acknowledge that kernel-specific training could be

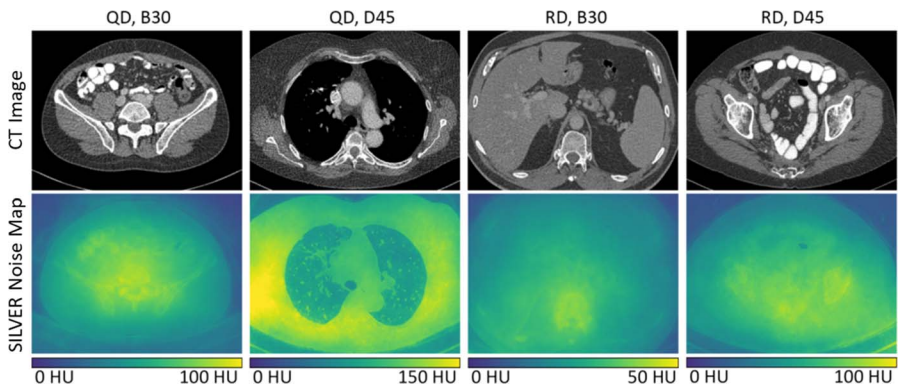

FIGURE 3. Representative patient CT images and corresponding SILVER noise map prediction, at 2 dose levels (RD and QD) and 2 kernels (B30: smooth, D45: medium sharp). Notice the large variety in noise levels and textures observed within the patient data set. Image selection was made to show performance within lungs, abdomen, and pelvis. This figure can be viewed online in color at www.jcat.org.

TABLE 2. Average Absolute Percent Error of SILVER Versus ROI SD Measurement for Preselected Regions in Patient CT Images (Heart, Aorta, Liver, Spleen, Fat)

	QD: Percent Error, %		RD: Percent Error, %	
Region	**B30**	**D45**	**B30**	**D45**
Heart	6 ± 4	4 ± 2	7 ± 7	4 ± 4
Aorta	7 ± 5	4 ± 3	9 ± 5	5 ± 3
Liver	4 ± 2	4 ± 3	5 ± 4	3 ± 2
Spleen	4 ± 4	3 ± 3	7 ± 5	4 ± 3
Fat	6 ± 4	4 ± 3	7 ± 8	4 ± 4

Error bars reflect the SD of percent error in 10 patient examinations.

a barrier for use in a clinical setting where multiple kernels are used. In future work, we plan to include multiple kernels within the SILVER training data set to improve generalizability and quantify performance when applied to different reconstruction kernels.

CONCLUSIONS

The investigated CNN-based technique was capable of accurately predicting the noise level directly from CT phantom and patient images. The SILVER noise map provided noise level estimation of nonuniform regions, which is unattainable using existing methods of noise quantification. This article provides an example of the potential benefit of using deep learning for patient-specific image quality assessment.

ACKNOWLEDGMENT

The authors wish to acknowledge Cynthia McCollough, PhD, the Mayo Clinic, the American Association of Physicists in Medicine, and the National Institute of Biomedical Imaging and Bioengineering (grants EB017095 and EB017185) for distributing the data used within this publication and Mr. Kevin Kimlinger for his assistance with manuscript preparation.

REFERENCES

1. Christianson O, Winslow J, Frush DP, et al. Automated technique to measure noise in clinical CT examinations. *Am J Roentgenol*. 2015;205:W93–W99.
2. Ahmad M, Jacobsen MC, Thomas MA, et al. A benchmark for automatic noise measurement in clinical computed tomography. *Med Phys*. 2021;48:640–647.
3. Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. *Med Phys*. 2017;44:2173–2184.
4. Anam C, Budi WS, Adi K, et al. Assessment of patient dose and noise level of clinical CT images: automated measurements. *J Radiol Prot*. 2019;39:783–793.
5. Chun M, Choi YH, Kim JH. Automated measurement of CT noise in patient images with a novel structure coherence feature. *Phys Med Biol*. 2015;60:9107–9122.
6. Tian X, Samei E. Accurate assessment and prediction of noise in clinical CT images. *Med Phys*. 2016;43:475–482.
7. Anam C, Arif I, Haryanto F, et al. An improved method of automated noise measurement system in CT images. *J Biomed Phys Eng*. 2021;11:163–174.
8. Ma J, Li S, He J, et al. Blind CT image quality assessment via deep learning strategy: initial study. In: Nishikawa RM, Samuelson FW, eds. *Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment*. Bellingham, WA: SPIE; 2018:44.
9. Imran AAZ, Pal D, Patel B, et al. Ssiqa: Multi-Task Learning for Non-Reference Ct Image Quality Assessment With Self-Supervised Noise Level Prediction. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 2021:1962–1965.
10. Zhang C, Gomez D, Li Y, et al. Patient-specific noise power spectrum measurement via generative adversarial networks. *SPIE Med Imaging Proc*. 2019;10948.
11. Pan X, Yu L. Image reconstruction with shift-variant filtration and its implication for noise and resolution properties in fan-beam computed tomography. *Med Phys*. 2003;30:590–600.
12. Wunderlich A, Noo F. Image covariance and lesion detectability in direct fan-beam x-ray computed tomography. *Phys Med Biol*. 2008;53:2471–2493.
13. Li Z, Yu L, Trzasko JD, et al. Adaptive nonlocal means filtering based on local noise level for CT denoising. *Med Phys*. 2014;41:011908.
14. Chen H, Zhang Y, Kalra MK, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36:2524–2535.
15. Huber N, Anderson T, Missert A, et al. Clinical evaluation of a phantom-based deep convolutional neural network for whole-body-low-dose and ultra-low-dose CT skeletal surveys. *Skeletal Radiol*. 2022;51:145–151.
16. McCollough CH, Bartley AC, Carter RE, 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*. 2017;44:e339–e352.
17. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation [published online May 18, 2015]. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science. Vol 9351. Cham, Switzerland: Springer; 2015. 10.1007/978-3-319-24574-4_28.
18. CT Clinical Innovation Center. Mayo_Grand_Challenge | Powered by Box. Box published August 25, 2021. Available at: https://aapm.app.box.com/s/eaw4jddb53keg1bptavvvd1sf4x3pe9h. Accessed February 17, 2022
19. Yu L, Shiung M, Jondal D, et al. Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. *J Comput Assist Tomogr*. 2012;36:477–487.
20. Horiuchi T, Yamamoto S, Murase K, et al. Development of DICOM image-based CT low dose simulator using fan-beam transform. *Technol Health Care*. 2013;21:441–454.
21. Alsaihati N, Solomon J, Samei E. Development and validation of a generic image-based noise addition software for simulating reduced dose computed tomography images using synthetic projections. *SPIE Med Imaging Proc*. 2022;12312U.
22. Huber NR, Missert AD, Yu L, et al. Evaluating a convolutional neural network noise reduction method when applied to CT images reconstructed differently than training data. *J Comput Assist Tomogr*. 2021;45:544–551.
23. Zeng R, Lin CY, Li Q, et al. Performance of a deep learning–based CT image denoising method: generalizability over dose, reconstruction kernel, and slice thickness. *Med Phys*. 2022;49:836–853.