CTFlow: Mitigating Effects of Computed Tomography Acquisition and Reconstruction with Normalizing Flows

Leihao Wei[1,2], Anil Yadav[2], and William Hsu[2](✉)

[1] Department of Electrical and Computer Engineering, Samueli School of Engineering, University of California, Los Angeles, CA 90095, USA
[2] Medical and Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA 90024, USA
whsu@mednet.ucla.edu

Abstract. Mitigating the effects of image appearance due to variations in computed tomography (CT) acquisition and reconstruction parameters is a challenging inverse problem. We present CTFlow, a normalizing flows-based method for harmonizing CT scans acquired and reconstructed using different doses and kernels to a target scan. Unlike existing state-of-the-art image harmonization approaches that only generate a single output, flow-based methods learn the explicit conditional density and output the entire spectrum of plausible reconstruction, reflecting the underlying uncertainty of the problem. We demonstrate how normalizing flows reduces variability in image quality and the performance of a machine learning algorithm for lung nodule detection. We evaluate the performance of CTFlow by 1) comparing it with other techniques on a denoising task using the AAPM-Mayo Clinical Low-Dose CT Grand Challenge dataset, and 2) demonstrating consistency in nodule detection performance across 186 real-world low-dose CT chest scans acquired at our institution. CTFlow performs better in the denoising task for both peak signal-to-noise ratio and perceptual quality metrics. Moreover, CTFlow produces more consistent predictions across all dose and kernel conditions than generative adversarial network (GAN)-based image harmonization on a lung nodule detection task. The code is available at https://github.com/hsu-lab/ctflow.

Keywords: Image harmonization · computed tomography · normalizing flows

1 Introduction

The increased availability of radiological data and rapid advances in medical image analysis has led to an exponential growth in prediction models that utilize features extracted from clinical imaging scans to detect and diagnose diseases and predict response to treatment [1–4]. However, variations in the acquisition and reconstruction of CT scans

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_39.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 413–422, 2023.
https://doi.org/10.1007/978-3-031-43990-2_39

result in quantitative image features with poor reproducibility [5, 6]. Several studies have demonstrated that differences in dose, slice thickness, reconstruction method, and reconstruction kernel negatively impact radiomic feature reproducibility. Predicting prediction model performance is confounded by how medical images are acquired and reconstructed [7–10]. Many studies have developed techniques to address sources of CT parameter variability [5, 11]. However, inverse problems such as recovering full radiation dose scans from lower dose scans are inherently ill-posed. A range of outputs may be possible when using image restoration algorithms, including potential artifacts that impact the performance of downstream algorithms. Like GANs or variational autoencoders, normalizing flows is a method for learning complex data representations but with an explicit ability to infer the output as a probability distribution and with the added benefit of more stable training [12]. While normalizing flows has shown success in image synthesis tasks for natural images [13], few studies have examined them in medical image harmonization tasks. Denker et al. employed a normalizing flow model conditioned on LDCT reconstruction by filtered backprojection to improve reconstruction quality from the raw sinogram data [14].

This paper presents CTFlow, which aims to utilize normalizing flows to harmonize variations in image appearance of CT scans by maximizing the explicit likelihood of a target condition (e.g., 100% dose, medium kernel, 1 mm slice thickness) given a CT scan that was acquired using different parameters (50% dose, sharp kernel, 1 mm slice thickness). Normalizing flow has two important advantages: 1) the translated low-dose CTs have minimal artifacts because the output is a maximum likelihood estimate that closely matches the target reference distribution, and 2) unlike GANs, which are susceptible to mode collapse, CTFlow can generate multiple solutions to reduce inference uncertainty. We demonstrate how CTFlow compares with current state-of-the-art methods for mitigating dose and reconstruction kernel differences. We evaluated using image quality metrics and a lung nodule detection task.

2 Methods

2.1 Datasets

This study used two unique datasets: (1) the UCLA low-dose chest CT dataset, a collection of 186 exams acquired using Siemens CT scanners at an equivalent dose of 2 mGy following an institutional review board-approved protocol. The raw projection data of scans were exported, and Poisson noise was introduced, as described in Zabic *et al.* [15], at levels equivalent to 10% of the original dose. Projection data were then reconstructed into an image size of 512 × 512 using three reconstruction kernels (smooth, medium, sharp) at 1.0 mm slice thickness. The dataset was split into 80 scans for training, 20 for validation, and 86 for testing. (2) AAPM-Mayo Clinic Low-Dose CT "Grand Challenge" dataset, a publicly available Grand Challenge dataset consisting of 5,936 abdominal CT images from 10 patient cases reconstructed at 1.0 mm slice thickness. Each case consists of a paired 100% "normal dose" scan and a simulated 25% "low dose" scan. Images from eight patient cases were used for training, and two cases were reserved for validation. All images were randomly cropped into patches of 128 × 128 pixels, generated from regions containing the body. This dataset was only used for evaluating image quality against other harmonization techniques.

2.2 Normalizing Flows

In this section, we describe the normalizing flows and modifications that were made to improve computational efficiency. Deterministic approaches to image translation (e.g., using a convolutional neural network) attempt to find a mapping function $y = g_\theta(x)$ that takes an input image x and outputs an image y that mimics the appearance of a target condition. For example, x could be an image acquired using a low dose protocol (e.g., 25% dose, smooth kernel), and y represents the image acquired at the target acquisition and reconstruction parameter (e.g., 100% dose, medium kernel). Flow-based image translation aims to approximate the density function $\prod_{y|x}(y|x, \theta)$ using maximum likelihood estimation.

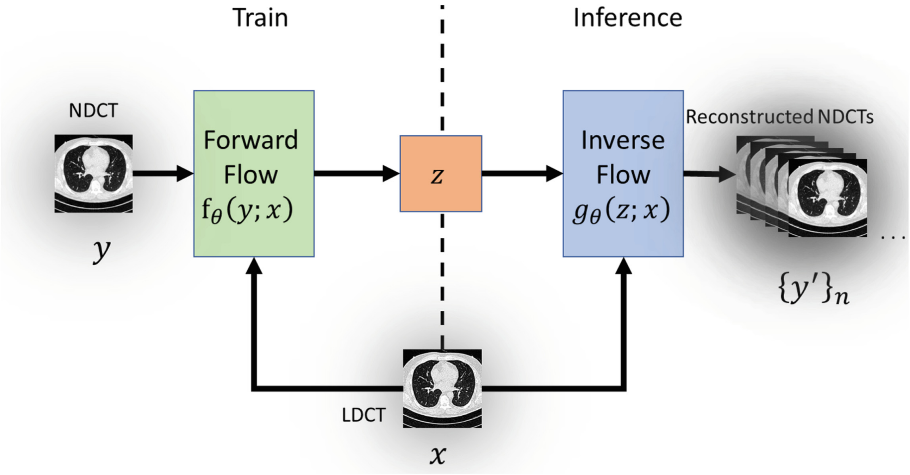


Fig. 1. Relationship between forward flow and inverse flow. LDCT: low (25%) dose computed tomography scan; NDCT: normal (100%) dose computed tomography scan.

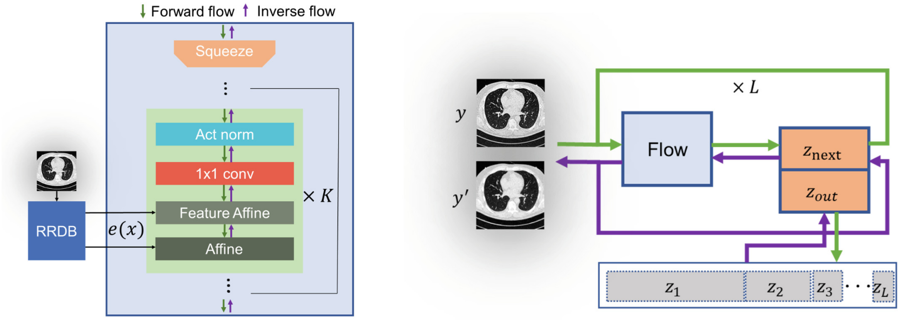
Figure 1 summarizes the normalizing flow approach. Normalizing flow gradually transforms an initial (Gaussian) density function $p_z(z)$ to a target distribution $\prod(y|x)$ using an invertible neural network $y = g_\theta(z; x) \leftrightarrow z = g_\theta^{-1}(y; x) = f_\theta(y; x)$, where g and$f$ are the decoding (inverse flow) and encoding (forward flow) functions, respectively. By the change of variables theorem, we can calculate the density function

$$\prod\nolimits_{y|x}(y|x, \theta) = p_z(f_\theta(y; x)) \left| det \frac{df_\theta(y; x)}{dy} \right| \tag{1}$$

which can be trained by maximizing the log-likelihood. In practice, a multilayer flow operation is preferred because a single-layer flow cannot represent complex non-linear relationships within the data. f is decomposed into a series of invertible neural network layers h^n where $h^n = f_\theta^n(h^{n-1}; e(x))$, n represents the number of layers, and $e(x)$ represents a deep convolutional neural network that extracts salient feature maps of x upon which the flow layers are conditioned. For an N-layer flow model, the objective is to maximize

$$\hat{\theta} = \underset{\theta}{\text{argmax}} logp_z(z) + \sum_{n=1}^{N} log\left| det \frac{df_\theta^n(h^{n-1}; e(x))}{dh^{n-1}} \right| \tag{2}$$

Once the training is complete, the decoding function $g_\theta(z; x)$ is applied using random latent variable z, which is drawn from the independent and identically distributed Gaussian density function. The use of z allows us to generate a range of possible restored images $y\prime$, conditioned on the same input image x.



(a) (b)

Fig. 2. (a) Flow module. RRDB: Residual in Residual Dense Block (b) Multiscale architecture

Flow Layers. Flow layers must meet two requirements: 1) be invertible and 2) be a tractable Jacobian determinant. To compute the second term in Eq. 2, we apply the triangulation trick developed by Dinh *et al.* [16] We use affine coupling layers with a conditional variable. We first equally split the channels into $\left(h_1^n, h_2^n\right)$ and apply an affine transformation on h_2^n while keeping an identity transform on h_1^n. We apply scale and shift factor computed by a shallow convolutional neural network (CNN) given h_1^n in spatial coordinates i, j to compute the n + 1 layer flow of h_2^{n+1}. Finally, we concatenated the splitting components back to obtain the next layer flow $h^{n+1} = concat\left(h_1^{n+1}, h_2^{n+1}\right)$ Thus, by definition, Jacobian of h^{n+1} is a lower triangular matrix. Figure 2a depicts the components of the flow module, which are described below:

- **Activation normalization:** A channel-wise batch normalization [17] was applied, yielding an output with zero mean and unit variance.
- **Invertible 1 x 1 conv:** Following the approach in [18], we utilized a learnable 1x1 convolution $h_{i,j}^n = Wh_{i,j}^{n-1}$ where W is a square matrix with dimension $c \times c$ (c is the number of channels). Each spatial element i, j in h is multiplied by this 1x1 convolution matrix. The log determinant is computed using PLU factorization.

- **Feature conditional affine.** We compute the scale and shift factor from $e(x)$ again using a shallow CNN to apply the n-th layer flow transformation h. The motivation is to impose a relationship between feature maps extracted $e(x)$ and activation maps h.

The deep convolutional neural network extractor $e(x)$ is based on Residual-in-Residual Dense Blocks (RRDB) [19]. This network contains 14 RRDB blocks and is our feature extractor for low-dose images. The RRDB network was trained using L_1 loss for 60k iterations. The batch size was 16 and the learning rate was set to 2e−4. The Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.99$. After training, all layers of RRDB were frozen and used only for feature extraction. Feature maps were derived from 2, 6, 10, 14 block outputs. Afterward, the outputs of each block were concatenated into $e(x)$.

Multiscale Architecture. Since the flow approach is invertible, input x and latent space vector z must have the same dimensions. However, in most cases, $\prod_{y|x}(y|x, \theta)$ is a low-dimensional manifold in high-dimensional input space. Computation is inefficient when a flow model is imposed with a higher dimensionality than the dimension of true latent space. Given the multiscale architecture in RealNVP, we can simplify the model and improve the density estimation at multiple levels. The overall multiscale architecture is depicted in Fig. 2b, where we equally divide each output z into (z_{out}, z_{next}), while recursively feeding z_{next} to the next level. Once all levels have been reached, z_{out} is outputted, representing the maximum log-likelihood estimation.

Network Training. We trained CTFlow using a batch size of 16 and 50k iterations. The learning rate was set to 1e−4 and halved at 50%, 75%, 90%, and 95% of the total training steps. A negative log-likelihood loss was used.

2.3 Experiments

We conducted two experiments to evaluate CTFlow: image quality metrics and impact on the performance of a lung nodule computer-aided detection (CADe) algorithm.

Image Quality. Using the Grand Challenge dataset, we assessed image quality and compared it with other previously published low-dose CT denoising techniques. We computed image quality metrics using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [20]. Our comparison was conducted using adversarial-based approaches (WGAN using mean squared error loss, WGAN using perceptual loss, and a 3D spectral-norm GAN called SNGAN [21–23], previously developed in our group), a convolutional neural network-based approach (SRResNet) [24], and a denoising algorithm based on collaborative filtering Block-matching and 3D filtering (BM3D) [25].

Nodule Detection. We evaluated the ability of CTFlow to harmonize differences in reconstruction kernels and their effect on the performance of a lung nodule detection algorithm. Our CADe system was based on the RetinaNet model, a composite model comprised of a backbone network called feature pyramid net and two subnetworks responsible for object classification with bounding box regression. The model was trained and validated on the LIDC-IDRI dataset, a public de-identified dataset of diagnostic and low-dose CT scans with annotations from four experienced thoracic radiologists. As part of the training process, we only considered nodules annotated by at least three readers in the LIDC dataset. A total of 7,607 slices (with 4,234 nodule annotations) were used for training and 2,323 slices (with 1,454 nodule annotations) for testing in a single train-test split. A bounding box was then created around the union of all the annotator contours to serve as the reference for the detection model. After training for 200 epochs with Focal loss and Adam optimizer, the model achieved an average precision (AP@0.5) of 0.62 on the validation set.

We hypothesized that the CTFlow models should yield better consistency in lung nodule detection performance compared with not normalizing or other state-of-the-art methods. As a comparison, we trained a 3D SNGAN model using the same training and validation set as CTFlow to perform the same task. We trained three separate CTFlow and SNGAN models to map scans reconstructed using smooth, medium, or sharp kernels to a reference condition. We computed the F1 score (the harmonic mean of the CADe algorithm's precision and recall) when executing the model on the CTFlow and SNGAN normalized scans. We then determined the Concordance Correlation Coefficient [26] on the F1 scores, comparing the F1 score of the model when executed on the normalized scan to when executed on the reference scan.

3 Results

3.1 Network Training

On the Grand Challenge dataset, CTFlow took 3 days to train on an NVIDIA RTX 8000 GPU. The peak GPU memory usage was 39 GB. Unlike GANs that required two loss functions, our network was optimized with only one loss function. The negative log-likelihood loss was stable and decreased monotonically.

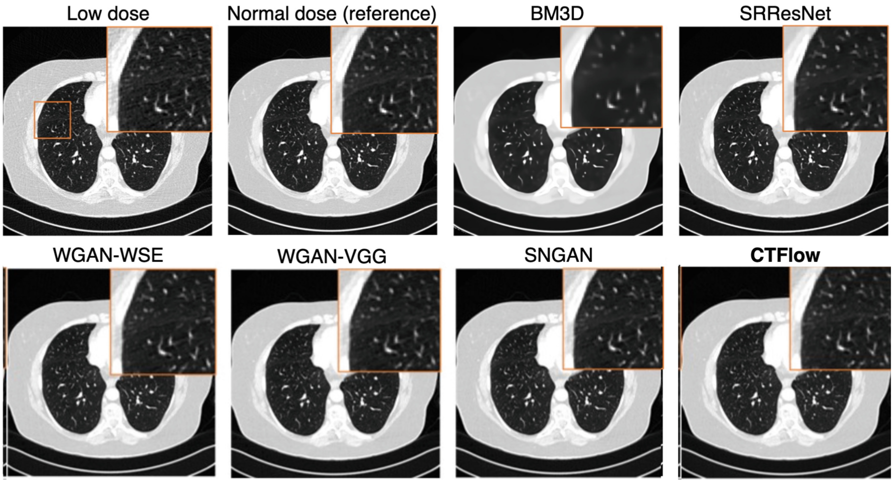


Fig. 3. Visual comparison of Grand Challenge dataset results.

3.2 Image Quality

Table 1 summarizes the results for image quality metrics, while Fig. 3 depicts the same representative slice outputted by each method. While BM3D and SRResNet generated the highest PSNR and SSIM, the images were overly smooth and lacked high-frequency components. Important texture details were lost in the restoration, which may negatively impact downstream tasks (e.g., radiologist interpretation, CAD algorithm performance) that rely on maintaining texture features to characterize lesions. CTFlow achieved 6% better perceptual quality than SNGAN.

Table 1. Image quality metrics generated from the Grand Challenge dataset validation images.

	PSNR	*SSIM*	*LPIPS*
BM3D	**32.81**	0.847	0.175
SRResNet	31.89	**0.891**	0.087
WGAN-MSE	31.32	0.864	0.036
WGAN-VGG	31.37	0.869	0.035
SNGAN	31.28	0.865	0.035
CTFlow	31.50	0.863	**0.032**

3.3 Nodule Detection

Table 2 summarizes the CCC values for each kernel pair. McBride [27] suggested the following guidelines for interpreting Lin's concordance correlation coefficient. Poor: <0:9; moderate: 0.90 to 0.95; substantial: 0.95 to 0.99; perfect: >0.99 and above. CTFlow achieved CCC scores within the "perfect" range when assessing the agreement in F1 scores when given images reconstructed using varying kernels.

Table 2. Consistency in lung nodule detection performance measured by CCC scores for three pairwise kernel combinations.

	Smooth to Medium	*Medium to Sharp*	*Smooth to Sharp*	*Mean*
SNGAN	0.853	0.844	0.941	0.879
CTFlow	**0.991**	**0.997**	**0.991**	**0.993**

4 Conclusion

We developed CTFlow, a normalizing flows approach to mitigating variations in CT scans. We demonstrated that CTFlow achieved consistent performance across image quality metrics, yielding the best perceptual quality score. Moreover, CTFlow was better than a GAN-based method in maintaining consistent lung nodule detection performance. Compared to generative models, the normalizing flows approach offers exact and efficient likelihood computation and generates diverse outputs that are closer to the target distribution.

We note several limitations of this work. In our evaluations, we trained separate CTFlow and comparison models for each mapping (e.g., transforming a 'smooth' kernel to a 'medium' kernel scan), allowing us to troubleshoot models more easily. A single model conditioned on different doses and kernels would be more practical. Also, CTFlow depends on tuning a variance parameter; better PSNR and SSIM may have been achieved with the optimization of this parameter. Finally, this study focused on mitigating the effect of a single CT parameter, either dose (in image quality) or kernel (in nodule detection). In the real world, multiple CT parameters interact (dose and kernel); these more complex interactions are being investigated as part of future work.

One underexplored area of normalizing flow is its ability to generate the full distribution of possible outputs. Using this information, we can estimate where high uncertainty exists in the model output, providing information to downstream image processing steps, such as segmentation, object detection, and classification. For example, Chan *et al.* [28] applied an approximate Bayesian inference scheme based on posterior regularization to improve uncertainty quantification on covariate-shifted data sets, resulting in improved prognostic models for prostate cancer. Investigating how this uncertainty can be incorporated into downstream tasks, such as our lung nodule CADe algorithm, is also part of future work.

Acknowledgments. This work was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under awards R56 EB031993 and R01 EB031993. The authors thank John M. Hoffman, Nastaran Emaminejad, and Michael McNitt-Gray for providing access to the UCLA low-dose CT dataset. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Sala, E., et al.: Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin. Radiol. **72**(1), 3 (2017). https://doi.org/10.1016/j.crad.2016.09.013
2. Fave, X., et al.: Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci. Rep. **7**(1), 588 (2017). https://doi.org/10.1038/s41598-017-00665-z
3. Aerts, H.J.: The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol. **2**(12), 1636–1642 (2016). https://doi.org/10.1001/jamaoncol.2016.2631
4. Aerts, H.J., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**(1), 4006 (2014). https://doi.org/10.1038/ncomms5006
5. Traverso, A., Wee, L., Dekker, A., Gillies, R.: repeatability and reproducibility of radiomic features: a systematic review. Int. J. Radiat. Oncol. Biol. Phys. **102**(4), 1143–1158 (2018). https://doi.org/10.1016/j.ijrobp.2018.05.053
6. Mackin, D., et al.: Measuring computed tomography scanner variability of radiomics features. Invest. Radiol. **50**(11), 757–765 (2015). https://doi.org/10.1097/RLI.0000000000000180

7. Lu, L., Ehmke, R.C., Schwartz, L.H., Zhao, B.: Assessing agreement between radiomic features computed for multiple CT imaging settings. PLoS ONE **11**(12), e0166550 (2016). https://doi.org/10.1371/journal.pone.0166550
8. Zhao, B., et al.: Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci. Rep. **6**, 23428 (2016). https://doi.org/10.1038/srep23428
9. Kalpathy-Cramer, J., et al.: Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography. **2**(4), 430–437 (2016). https://doi.org/10.18383/j.tom.2016.00235
10. Lo, P., Young, S., Kim, H.J., Brown, M.S., McNitt-Gray, M.F.: Variability in CT lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. Med. Phys. **43**(8), 4854 (2016). https://doi.org/10.1118/1.4954845
11. Nan, Y., et al.: Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. Inf. Fusion **82**, 99–122 (2022). https://doi.org/10.1016/j.inffus.2022.01.001
12. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using Real NVP (2016)
13. Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R.: SRFlow: learning the super-resolution space with normalizing flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 715–732. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_42
14. Denker, A., Schmidt, M., Leuschner, J., Maass, P., Behrmann, J.: Conditional normalizing flows for low-dose computed tomography image reconstruction (2020)
15. Zabic, S., Wang, Q., Morton, T., Brown, K.M.: A low dose simulation tool for CT systems with energy integrating detectors. Med. Phys. **40**(3), 031102 (2013). https://doi.org/10.1118/1.4789628
16. Dinh, L., Krueger, D., Bengio, Y.: Nice: non-linear independent components estimation. arXiv preprint arXiv:14108516 (2014)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–56. PMLR (2015)
18. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
19. Wang, X., et al.: EsrGAN: enhanced super-resolution generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 63–79. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_5
20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
21. Wolterink, J.M., Leiner, T., Viergever, M.A., Isgum, I.: Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans. Med. Imaging **36**(12), 2536–2545 (2017). https://doi.org/10.1109/TMI.2017.2708987
22. Yang, Q., et al.: Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging. **37**(6), 1348–1357 (2018). https://doi.org/10.1109/TMI.2018.2827462
23. Wei, L., Lin, Y., Hsu, W.: Using a generative adversarial network for CT normalization and its impact on radiomic features. In: IEEE International Symposium on Biomedical Imaging. Iowa City, IA (2020)
24. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution (2017)
25. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans Image Process. **16**(8), 2080–2095 (2007). https://doi.org/10.1109/tip.2007.901238

26. Lawrence, I., Lin, K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics, 255–268 (1989)
27. McBride, G.: A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA client report: HAM2005-062, p. 62 (2005)
28. Chan, A., Alaa, A., Qian, Z., Van Der Schaar, M.: Unlabelled data improves Bayesian uncertainty calibration under covariate shift. In: International Conference on Machine Learning, pp. 1392–402. PMLR (2020)