Inter-slice Consistency for Unpaired Low-Dose CT Denoising Using Boosted Contrastive Learning

Jie Jing[1], Tao Wang[1], Hui Yu[1], Zexin Lu[1], and Yi Zhang[2(✉)]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
[2] School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China
yzhang@scu.edu.cn

Abstract. The research field of low-dose computed tomography (LDCT) denoising is primarily dominated by supervised learning-based approaches, which necessitate the accurate registration of LDCT images and the corresponding NDCT images. However, since obtaining well-paired data is not always feasible in real clinical practice, unsupervised methods have become increasingly popular for LDCT denoising. One commonly used method is CycleGAN, but the training processing of CycleGAN is memory-intensive and mode collapse may occur. To address these limitations, we propose a novel unsupervised method based on boosted contrastive learning (BCL), which requires only a single generator. Furthermore, the constraints of computational power and memory capacity often force most existing approaches to focus solely on individual slices, leading to inconsistency in the results between consecutive slices. Our proposed BCL-based model integrates inter-slice features while maintaining the computational cost at an acceptable level comparable to most slice-based methods. Two modifications are introduced to the original contrastive learning method, including weight optimization for positive-negative pairs and imposing constraints on difference invariants. Experiments demonstrate that our method outperforms existing several state-of-the-art supervised and unsupervised methods in both qualitative and quantitative metrics.

Keywords: Low-dose computed tomography · unsupervised learning · image denoising · machine learning

1 Introduction

Computed tomography (CT) is a common tool for medical diagnosis but increased usage has led to concerns about the possible risks caused by excessive radiation exposure. The well-known ALARA (as low as reasonably achievable)

This work was supported in part by the National Natural Science Foundation of China under Grant 62271335; in part by the Sichuan Science and Technology Program under Grant 2021JDJQ0024; and in part by the Sichuan University "From 0 to 1" Innovative Research Program under Grant 2022SCUH0016.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 238–247, 2023.
https://doi.org/10.1007/978-3-031-43907-0_23

[20] principle is widely adopted to reduce exposure based on the strategies such as sparse sampling and tube flux reduction. However, reducing radiation dose will degrade the imaging quality and then inevitably jeopardize the subsequent diagnoses. Various algorithms have been developed to address this issue, which can be roughly categorized into sinogram domain filtration [16], iterative reconstruction [2,9], and image post-processing [1,8,14].

Recently, deep learning (DL) has been introduced for low-dose computed tomography (LDCT) image restoration. The utilization of convolutional neural networks (CNNs) for image super-resolution, as described in [7], outperformed most conventional techniques. As a result, it was subsequently employed for LDCT in [6]. The DIP [21] method is an unsupervised image restoration technique that leverages the inherent ability of untrained networks to capture image statistics. Other methods that do not require clean images are also used in this field [11,23]. Various network architectures have been proposed, such as RED [5] and MAP-NN [18]. The choice of loss function also significantly affects model performance. Perceptual loss [12] based on the pretrained VGG [19] was proposed to mitigate over-smoothing caused by MSE. Most DL techniques for LDCT denoising are supervised models, but unsupervised learning frameworks which do not need paired data for training like GANs [3,10,26], Invertible Network [4] and CUT [25] have also been applied for LDCT [13,15,24].

This study presents a novel unsupervised framework for denoising low-dose CT (LDCT) images, which utilizes contrastive learning (CL) and doesn't require paired data. Our approach possesses three major contributions as follows: Firstly, We discard the use of CycleGAN that most unpaired frameworks employ, instead adopting contrastive learning to design the training framework. As a result, the training process becomes more stable and imposes a lesser computational burden. Secondly, our approach can adapt to almost all end-to-end image translation neural networks, demonstrating excellent flexibility. Lastly, the proposed inter-slice consistency loss makes our model generates stable output quality across slices, in contrast to most slice based methods that exhibit inter-slice instability. Our model outperforms almost all other models in this regard, making it the superior option for LDCT denoising. Further experimental data about this point will be presented in this paper.

2 Method

LDCT image denoising can be expressed as a noise reduction problem in the image domain as $\hat{x} = f(x)$, where $\hat{x}$ and x denote the denoised output and corresponding LDCT image. f represents the denoising function. Rather than directly denoising LDCT images, an encoder-decoder model is used to extract important features from the LDCT images and predict corresponding NDCT images. Most CNN-based LDCT denoising models are based on supervised learning and require both the LDCT and its perfectly paired NDCT images to learn f. However, it is infeasible in real clinical practice. Currently, some unsupervised models, including CUT and CycleGAN, relax the constraint on requiring paired data for training. Instead, these models can be trained with unpaired data.

2.1 Contrastive Learning for Unpaired Data

The task of LDCT image denoising can be viewed as an image translation process from LDCT to NDCT. CUT provides a powerful framework for training a model to complete image-to-image translation tasks. The main concept behind CUT is to use contrastive learning for enhanced feature extraction aided by an adversarial loss.

The key principle of contrastive learning is to create positive and negative pairs of samples, in order to help the model gain strong feature representation ability. The loss of contrastive learning can be formulated as:

$$l(v, v^+, v^-) = -log[\frac{exp(v \cdot v^+/\tau)}{exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} exp(v \cdot v_n^-/\tau)}], \quad (1)$$

where v, v^+, v^- denote the anchors, positive and negative pairs, respectively. N is the number of negative pairs. τ is the temperature factor which is set to 0.07 in this paper. The generator G we used contains two parts, an encoder E and a decoder. A simple MLP H is used to module the features extracted from the encoder. The total loss of CUT for image translation is defined as:

$$L = L_{GAN}(G, D, X, Y) + \lambda_1 L_{PatchNCE}(G, H, X) + \lambda_2 L_{PatchNCE}(G, H, Y), \quad (2)$$

where D denotes the discriminator. X represents the input images, for which $L_{PatchNCE}(G, H, X)$ utilizes contrastive learning in the source domain (represented by noisy images). Y indicates the images in the target domain, which means NDCT images in this paper. $L_{PatchNCE}(G, H, Y)$ employs contrastive learning in this target domain. As noted in a previous study [25], this component plays a similar role as the identity loss in CycleGAN. In this work, λ_1 and λ_2 are both set to 1. Since CT images are three-dimensional data, we can identify more negative pairs between different slices. The strategy about how we design positive and negative pairs for our proposed model is illustrated in Fig. 1.

As shown in Fig. 1, we select two negative patches from the same slice as the anchor, as well as one from the previous slice and the other from the next slice. It is important to note that these patches are not adjacent, since neighbored slices are nearly identical. Similar to most contrastive learning methods, we use cosine similarity to compute the feature similarity.

2.2 Contrastive Learning for Inter-slice Consistency

Due to various constraints, most denoising methods for LDCT can only perform on the slice plane, resulting in detail loss among different slices. While 3D models can mitigate this issue to a certain degree, they require significant computational costs and are prone to model collapse during training, leading to a long training time. Additionally, most methods are unable to maintain structural consistency between slices with certain structures (e.g., bronchi and vessels) appearing continuously across several adjacent slices.

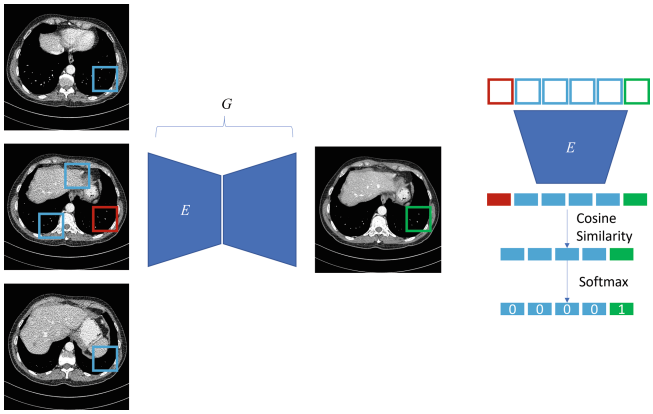


Fig. 1. The construction of training sample pairs for contrastive learning. The generator G is composed of the encoder E and the decoder. The anchor is represented by the red box, while the negative patches are indicated by blue boxes. Two of these negative patches come from the same slice but have different locations, while the other two come from different slices but have the same locations. The green box represents the positive patch, which is located in the same position as the anchor shown in the generated image. (Color figure online)

To address this issue, we design an inter-slice consistency loss based on contrastive learning. This approach helps to maintain structural consistency between slices, and then improve the overall denoising performance.

As illustrated in Fig. 2, we begin by randomly selecting the same patch from both the input (LDCT) and the generated denoised result. These patches are passed through the encoder E, allowing us to obtain the feature representation for each patch. Next, we perform a feature subtraction of each inter-slice pair. The output can be interpreted as the feature difference between slices. We assume that the feature difference between the same pair of slices should be similar, which is formulated as follows:

$$H(E(P(X_t)))-H(E(P(X_{t+1}))) = H(E(P(G(X_t))))-H(E(P(G(X_{t+1})))), \quad (3)$$

where P denotes the patch selection function. A good denoising generator can minimize the feature difference between similar slices while maximizing the feature difference between different slices. By utilizing contrastive learning, we can treat the former condition as a positive pair and the latter as a negative pair. After computing the cosine similarity of the pairs, a softmax operation is applied to assign 1 to the positive pairs and 0 to the negative pairs.

Compared to the original contrastive learning, which focuses on patch pairs, we apply this technique to measure feature differences, which stabilizes the features and improves the consistency between slices.

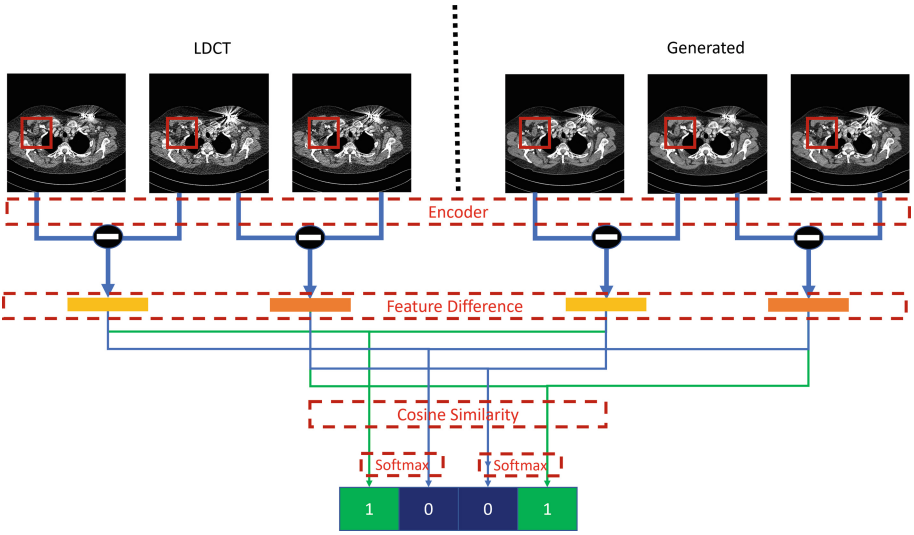


Fig. 2. Contrastive learning ultilized for stablizing inter-slice features. Patches are extracted from the same location in three consecutive slices.

2.3 Boosted Contrastive Learning

Original contrastive Learning approaches treat every positive and negative pair equally. However, in CT images, some patches may be very similar to others (e.g., patches from the same organ), while others may be completely different. Therefore, assigning the same weight to different pairs may not be appropriate. [25] demonstrated that fine-tuning the weights between pairs can significantly improve the performance of contrastive learning.

For our inter-slice consistency loss, only one positive and negative pair can be generated at a time, making it unnecessary to apply reweighting. However, we include additional negative pairs in the patchNCE loss for unpaired translation, making reweighting between pairs more critical than in the original CUT model. As a result, Eq. 1 is updated as follows:

$$l(v, v^+, v^-) = -log[\frac{exp(v \cdot v^+/\tau)}{exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} w_n exp(v \cdot v^-/\tau)}], \tag{4}$$

where w stands for a weight factor for each negative patch.

According to [25], using "easy weighting" is more effective for unpaired tasks, which involves assigning higher weights to easy negative samples (i.e., samples that are easy to distinguish from the anchor). This finding contradicts most people's intuition. Nonetheless, we have demonstrated that their discovery is accurate in our specific scenario. The reweighting approach we have employed is defined as follows:

$$w_n = \frac{exp(1 - v \cdot v_n^-)/\tau)}{\sum_{\substack{j=1 \\ j \neq n}}^{N} exp((1 - v \cdot v_j^-)/\tau)}. \tag{5}$$

In summary, the less similar two patches are, the easier they can be distinguished, the more weight the pair is given for learning purposes.

3 Experiments

3.1 Dataset and Training Details

While our method only requires unpaired data for training, many of the compared methods rely on paired NDCT. We utilized the dataset provided by the Mayo Clinic called "NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge" [17], which offers paired LD-NDCT images.

The model parameters were initialized using a random Gaussian distribution with zero-mean and standard deviation of 10^{-2}. The learning rate for the optimizer was set to 10^{-4} and halved every 5 epochs for 20 epochs total. The experiments were conducted in Python on a server with an RTX 3090 GPU. Two metrics, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [22], were employed to quantitatively evaluate the image quality. The image data from five individuals were used as the training set and the data from other two individuals were used for the test set.

3.2 Comparison of Different Methods

To demonstrate the denoising performance of our model, we conducted experiments to compare our method with various types of denoising methods including unsupervised denoising methods that only use LDCT data, fully supervised methods that use perfectly registered LDCT and NDCT pairs, and semi-supervised methods, including CycleGAN and CUT, which utilize unpaired data. A representative slice processed by different methods is shown in Fig. 3. The window center is set to 40 and the window width is set to 400.

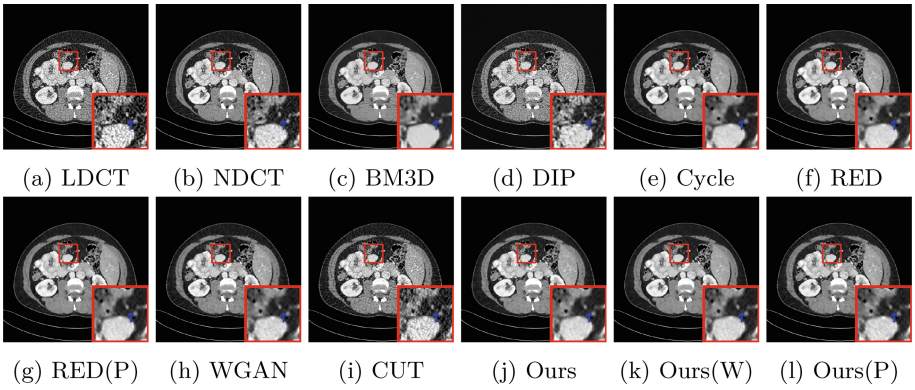
Our framework is flexible and can work with different autoencoder frameworks. In our experiments, the well-known residual encoder-decoder network (RED) was adopted as our network backbone.

The quantitative results and computational costs of unsupervised methods are presented in Table 1. It can be seen that our method produces promising denoising results, with obvious numerical improvements compared to other unsupervised and semi-supervised methods.

As shown in Table 2, our score is very close to our backbone model when trained fully supervised. Our model even got higher PSNR value.

Moreover, our framework is lightweight, which has a similar model scale to RED. It's worth noting that adding perceptual loss to our model will decrease the PSNR result, and it is consistent with the previous studies that perceptual loss may maintain more details but decrease the MSE-based metric, such as PSNR.

Furthermore, the reweighting mechanism demonstrates its effectiveness in improving our model's results. The improvement by introducing the reweighting mechanism can be easily noticed.



(a) LDCT (b) NDCT (c) BM3D (d) DIP (e) Cycle (f) RED

(g) RED(P) (h) WGAN (i) CUT (j) Ours (k) Ours(W) (l) Ours(P)

Fig. 3. Methods comparison. DIP and BM3D are fully unsupervised, RED and WGAN models will require paired dataset, CycleGAN("Cycle" in figure), CUT and ours will use unpaired dataset. "(P)" means perceptual loss is added. "(W)" means proposed re-weight mechanism is applied. (Color figure online)

Table 1. Metrics comparison for unsupervised methods. "(P)" means perceptual loss is added. "(W)" means proposed re-weight mechanism is applied.

Metrics	DIP	BM3D	CycleGAN	CUT	Ours	Ours(W)	Ours(P)	LDCT
PSNR	26.79	26.64	28.82	28.15	28.88	**29.09**	28.81	22.33
SSIM	0.86	0.82	**0.91**	0.86	0.90	**0.91**	**0.91**	0.63
MACs(G)	**75.64**	NaN	1576.09	496.87	521.36	521.36	521.36	NaN
Params(M)	2.18	NaN	7.58	3.82	**1.92**	**1.92**	**1.92**	NaN

3.3 Line Plot over Slices

Although our method may only be competitive with supervised methods, we are able to demonstrate the effectiveness of our proposed inter-slice consistency loss. The line plot in Fig. 4 shows the pixel values at point (200, 300) across different slices.

In Fig. 4, it can be observed that our method effectively preserves the inter-slice consistency of features, which is clinically important for maintaining the structural consistency of the entire volume. Although the supervised model achieves a similar overall score to our model, the results across slices of our model are closer to the ground truth (GT), especially when pixel value changes dramatically.

Table 2. Metrics comparison for supervised methods. "(P)" means perceptual loss is added. "(W)" means proposed re-weight mechanism is applied.

Metrics	RED(MSE)	RED(P)	WGAN	Ours(W)	LDCT
PSNR	29.06	28.74	27.75	**29.09**	22.33
SSIM	0.92	**0.92**	0.89	0.91	0.63
MACs(G)	**462.53**	**462.53**	626.89	521.36	NaN
Params(M)	**1.85**	**1.85**	2.52	1.92	NaN

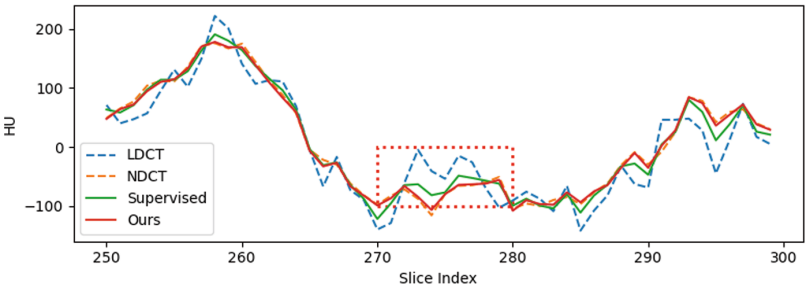


Fig. 4. Inter slice HU value line plot.

3.4 Discussion

Our method achieves competitive results and obtains the highest PSNR value in all the methods with unpaired samples. Although we cannot surpass supervised methods in terms of some metrics, our method produces promising results across consecutive slices that are more consistent and closer to the GT.

4 Conclusion

In this paper, we introduce a novel low-dose CT denoising model. The primary motivation for this work is based on the fact that most CNN-based denoising models require paired LD-NDCT images, while we usually can access unpaired CT data in clinical practice. Furthermore, many existing methods using unpaired samples require extensive computational costs, which can be prohibitive for clinical use. In addition, most existing methods focus on a single slice, which results in inconsistent results across consecutive slices. To overcome these limitations, we propose a novel unsupervised method based on contrastive learning that only requires a single generator. We also apply modifications to the original contrastive learning method to achieve SOTA denoising results using relatively a low computational cost.

Our experiments demonstrate that our method outperforms existing SOTA supervised, semi-supervised, and unsupervised methods in both qualitative and quantitative measures. Importantly, our framework does not require paired training data and is more adaptable for clinical use.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Beister, M., Kolditz, D., Kalender, W.A.: Iterative reconstruction methods in X-ray CT. Phys. Med. **28**(2), 94–108 (2012)
3. Bera, S., Biswas, P.K.: Axial consistent memory GAN with interslice consistency loss for low dose computed tomography image denoising. IEEE Trans. Radiation Plasma Med. Sci. (2023)
4. Bera, S., Biswas, P.K.: Self supervised low dose computed tomography image denoising using invertible network exploiting inter slice congruence. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5614–5623 (2023)
5. Chen, H., et al.: Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans. Med. Imaging **36**(12), 2524–2535 (2017)
6. Chen, H., Zet al.: Low-dose CT denoising with convolutional neural network. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 143–146. IEEE (2017)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
8. Feruglio, P.F., Vinegoni, C., Gros, J., Sbarbati, A., Weissleder, R.: Block matching 3D random noise filtering for absorption optical projection tomography. Phys. Med. Biol. **55**(18), 5401 (2010)
9. Geyer, L.L., et al.: State of the art: iterative CT reconstruction techniques. Radiology **276**(2), 339–357 (2015)
10. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
11. Jing, J., et al.: Training low dose CT denoising network without high quality reference data. Phys. Med. Biol. **67**(8), 084002 (2022)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
13. Jung, C., Lee, J., You, S., Ye, J.C.: Patch-wise deep metric learning for unsupervised low-dose ct denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 634–643. Springer (2022). https://doi.org/10.1007/978-3-031-16446-0_60
14. Kang, D., et al.: Image denoising of low-radiation dose coronary CT angiography by an adaptive block-matching 3D algorithm. In: Medical Imaging 2013: Image Processing. vol. 8669, pp. 86692G. International Society for Optics and Photonics (2013)
15. Li, Z., Huang, J., Yu, L., Chi, Y., Jin, M.: Low-dose CT image denoising using cycle-consistent adversarial networks. In: 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1–3 (2019). https://doi.org/10.1109/NSS/MIC42101.2019.9059965
16. Manduca, A., et al.: Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. Med. Phys. **36**(11), 4911–4919 (2009)
17. Moen, T.R., et al.: Low-dose CT image and projection dataset: . Med. Phys. **48**, 902–911 (2021)

18. Shan, H., et al.: Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nature Mach. Intell. **1**(6), 269–276 (2019)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Smith-Bindman, R., et al.: Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch. Intern. Med. **169**(22), 2078–2086 (2009)
21. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Deep image prior. CoRR abs/1711.10925 (2017). https://arxiv.org/abs/1711.10925
22. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
23. Wu, D., Gong, K., Kim, K., Li, X., Li, Q.: Consensus neural network for medical imaging denoising with only noisy training samples. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 741–749. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_81
24. 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
25. Zhan, F., Zhang, J., Yu, Y., Wu, R., Lu, S.: Modulated contrast for versatile image synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18259–18269 (2022).https://doi.org/10.1109/CVPR52688.2022.01774
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)