HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

IEEE Trans Med Imaging. 2023 June ; 42(6): 1590–1602. doi:10.1109/TMI.2022.3231428.

Noise Suppression with Similarity-based Self-Supervised Deep Learning

Chuang Niu,

Mengzhou Li,

Fenglei Fan,

Weiwen Wu,

Xiaodong Guo,

Qing Lyu,

Ge Wang

Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

Image denoising is a prerequisite for downstream tasks in many fields. Low-dose and photon-counting computed tomography (CT) denoising can optimize diagnostic performance at minimized radiation dose. Supervised deep denoising methods are popular but require paired clean or noisy samples that are often unavailable in practice. Limited by the independent noise assumption, current self-supervised denoising methods cannot process correlated noises as in CT images. Here we propose the first-of-its-kind similarity-based self-supervised deep denoising approach, referred to as Noise2Sim, that works in a nonlocal and nonlinear fashion to suppress not only independent but also correlated noises. Theoretically, Noise2Sim is asymptotically equivalent to supervised learning methods under mild conditions. Experimentally, Nosie2Sim recovers intrinsic features from noisy low-dose CT and photon-counting CT images as effectively as or even better than supervised learning methods on practical datasets visually, quantitatively and statistically. Noise2Sim is a general self-supervised denoising approach and has great potential in diverse applications.

Keywords

Self-supervised image denoising; low-dose CT denoising; photon-counting CT denoising

I. Introduction

Computed tomography (CT) reconstructs cross-sectional or volumetric images from many X-ray projections taken at different angles, and is a widely used diagnostic tool around the world. As an emerging CT technology, photon-counting CT (PCCT) is being actively developed with major advantages including rich tissue contrast, high spatial resolution, low radiation dose, and tracer-enhanced K edge imaging [1]. In the medical CT field, radiation

niuc@rpi.edu.

dose must be minimized according to the ”As Low As Reasonably Achievable” (ALARA) guideline [2]. Not surprisingly, reduced radiation dose will degrade the CT image quality and affect the diagnostic performance. In particular, PCCT uses multiple energy windows and much reduced detector sizes. As a result, each line integral measurement can only be made with a significantly lower number of x-ray photons. Over the past decade, great efforts have been made on low-dose CT (LDCT) reconstruction, also known as LDCT denoising, which has now a boosted momentum due to the recent FDA-approval of the PCCT technology [1].

Image denoising is to recover signals hidden in a noisy background. Since noise is a statistical fluctuation governed by quantum mechanics, denoising is generally achieved by a mean/averaging operation. For example, local averaging methods include Gaussian smoothing [3], anisotropic filtering [4], [5], neighborhood filtering [6]–[8], and transform domain processing [9]. On the other hand, nonlocal averaging methods use various nonlocal means with Gaussian kernel based weighting [10] or via nonlocal collaborative filtering in a transform domain [11]. Impressively, the nonlocal methods usually outperform the local methods, as images usually consist of repeated features or patterns that can be leveraged over a field of view to recover signals coherently. Over the past several years, the area of image denoising has been dominated by deep convolutional neural networks (CNNs) [12]. Different from the traditional methods that directly denoise an image based on an explicit model, the deep learning approach optimizes a deep neural network using training data, and then uses the trained model to predict a denoised image, usually achieving better results with much less inference time than the traditional denoising methods.

The mainstream deep denoising methods [13] require paired and registered noise-clean images to train the networks, denoted as Noise2Clean. However, such noise-clean image pairs can be hardly obtained in real-world applications. Especially for CT denoising tasks [14]–[17], it is impractical to scan the same patient twice to obtain paired training samples in a clinical setting. Although simulation tools [18], [19] can generate such paired training samples, it is very challenging to simulate realistic noise for real patients in many cases. Just taking CT as an example, after a modern CT reconstruction process, image noises are determined by many factors including imaging parameters/protocols, reconstruction steps, and objects being imaging, and hence much more difficult to simulate than in the case of naïve FBP reconstruction. Also, photon-counting CT is emerging as the next generation of CT technology but noise simulation for photon-counting data is much more complicated than the Poisson-Gaussian model successfully used for current-integrating CT. To relax the requirement of paired noise-clean samples, Noise2Noise [20] was proposed to train a denoising network with paired noise-noise images that share the same content but are instantiated with independent noises. It has been proved that Noise2Noise-based training can optimize a network to approach the Noise2Clean quality under the assumption of zero-mean noise. Some recent methods [21]–[27] heuristically adapted Noise2Noise for LDCT denoising by simulating paired noisy samples. However, the required collection of paired noisy images is still expensive in many scenarios and subject to potential mismatches between simulated and real noisy pairs. On the other hand, another important direction is the use of unpaired clean images as the targets to train the denoising model based on the generative adversarial network (GAN) [17]. Indeed, for LDCT denoising

various excellent unpaired learning methods were proposed, such as weakly-supervised progressive denoising [28], Cycle-Free CycleGAN [29], IdentityGAN and GAN-CIRCLE [30], AdaIN-Based Tunable CycleGAN [31]. Recently, self-supervised or unsupervised learning methods without using any extra annotations have achieved great progress in various domains [32]–[36], and extensive efforts [37]–[46] were also made to develop self-supervised learning methods for image denoising. Along this direction, Noise2Void [37] and Noise2Self [38] were proposed to predict each center pixel from its local neighbors, achieving promising results using single noisy images under the assumption of *independent* noises among neighbor pixels. However, these unsupervised deep denoising methods cannot suppress correlated or structured noises. Since most noises, especially CT image noises, are correlated, a general unsupervised denoising approach is highly desirable to unleash the power of deep learning for effective suppression of both independent and correlated noises, which is the holy grail of denoising methods.

Similarity including symmetry is ubiquitous in the physical world, naturally reflected in the image domain, featured by repeated or recursively embedded structures, edges, textures, etc., and plays an important role in modern science, engineering and medicine [47]. However, there is currently no general approach to utilize similar features in images for deep learning-based denoising. As the first effort to utilize similarity for unsupervised deep denoising, the initial version of our Noise2Sim method was shared on arXiv [48]. Here we present our Noise2Sim approach as the general framework for similarity-based optimization of a deep denoising network using intrinsically-registered sub-images. Importantly, the theorem is proved that under mild conditions the deep denoising network trained with Noise2Sim is asymptotically equivalent (with respect to the limit of the number of training samples) to that trained in the supervised learning mode. Given the equivalency between Noise2Clean and Noise2Sim, the proposed unsupervised learning approach can, in principle, suppress both independent and correlated noises as effectively as the supervised deep denoising methods [15], [49].

The emphasis of this paper is on the general self-supervised deep denoising approach and its practical application in LDCT and PCCT imaging where image noises are correlated and practically paired samples are usually not available. Extensive experiments on 3D LDCT images and 4D PCCT images consistently show the superiority of Noise2Sim over both the traditional denoising methods and the state-of-the-art deep denoising networks. Note that in this study we only focus on suppressing LDCT noise in the image domain, and the dual-domain unrolling algorithms [50] are not compared. Furthermore, we demonstrated the effectiveness of Noise2Sim on denoising 2D images with independent noises in Appendix A2. A recent study has shown that Noise2Sim can be successfully applied to other domains as well [51]. The source code, pre-trained models, and all data used in this study have been made publicly available on the project page http://chuangniu.info/projects/noise2im/, where the Python package “noise2sim” can be easily installed through Pip and source.

II. Methods

A. Theoretical Framework

A noisy image $\boldsymbol{x}_i$ can be decomposed into two parts: $\boldsymbol{x}_i = \boldsymbol{s}_i + \boldsymbol{n}_i$, which is generated from the joint distribution $p(\boldsymbol{s}, \boldsymbol{n}) = p(\boldsymbol{s})p(\boldsymbol{n} \mid \boldsymbol{s})$, where $\boldsymbol{s}_i$ and $\boldsymbol{n}_i$ are the clean signal and associated noise respectively. Note that this additive notation is notation-wise convenient without loss of generality, since a given noise sample can be practically decomposed into a real signal plus a noise component although the intrinsic noise model is not additive, as long as the conditions for the Noise2Sim theorem are satisfied. A deep denoising method learns a network function to recover the clean signal $\boldsymbol{s}_i$ from the noisy signal $\boldsymbol{x}_i$, *i.e.*, $\boldsymbol{y}_i = f(\boldsymbol{x}_i; \boldsymbol{\theta})$, where f denotes the network function with a vector of parameters $\boldsymbol{\theta}$ to be optimized. In a supervised training process (Noise2Clean), each noisy image $\boldsymbol{x}_i$ is associated with the corresponding clean image $\boldsymbol{s}_i$ as the target. Let $\boldsymbol{\theta}_c$ be the network parameters optimized with paired noise-clean data, we have

$$\boldsymbol{\theta}_c = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N_c} \sum_{i=1}^{N_c} \left\| f(\boldsymbol{s}_i + \boldsymbol{n}_i; \boldsymbol{\theta}) - \boldsymbol{s}_i \right\|_2^2, \tag{1}$$

where N_c is the number of images, and the mean squared error (MSE) is used as the loss function [20], [37]. Among all denoising methods, the supervised deep denoising methods generally achieve the best results but they are handicapped when paired images are unavailable.

The ubiquitous similarity in the physical world leads to the well-known popularity of similar patches, slices, volumes and tensors embedded within and across high-dimensional images. In this study, patches in 2D images, slices/patches in 3D images, and volumes in 4D images are regarded as sub-images. The similarity between sub-images is measured by a general function $S(T(\boldsymbol{x}_i), T(\hat{\boldsymbol{x}}_i))$, where T is a transform of a sub-image, S is a metric to measure the similarity between two sub-images. In practice, T and S may take different forms, depending on domain-specific priors. In principle, it would be an ideal strategy to match every sub-image up with all possible similar sub-images. Based on this fact, we propose Noise2Sim, a general unsupervised denoising approach for training a deep network without collecting paired noisy or clean target data. The key idea behind Noise2Sim is to replace paired clean or noisy targets with the similar targets for training the denoising network, such that noises are suppressed to enhance signals faithfully. Given noisy high-dimensional images only, we first search and construct a set of similar sub-images, denoted by $\boldsymbol{x}_i = \boldsymbol{s}_i + \boldsymbol{n}_i$ and $\hat{\boldsymbol{x}}_i = \boldsymbol{s}_i + \boldsymbol{\delta}_i + \hat{\boldsymbol{n}}_i$, where $\boldsymbol{\delta}_i$ is the difference between the clean signal components in similar sub-images, and $\boldsymbol{n}_i$ and $\hat{\boldsymbol{n}}_i$ are two different noise realizations. The similar sub-images are searched in a nonlocal way; *e.g.*, searching patches/slices within the whole image and even across different images. Let $\boldsymbol{\theta}_s$ be the vector of network parameters to be optimized with the constructed similar data by minimizing the following loss function:

$$\boldsymbol{\theta}_s = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| f(\boldsymbol{s}_i + \boldsymbol{n}_i; \boldsymbol{\theta}) - (\boldsymbol{s}_i + \boldsymbol{\delta}_i + \hat{\boldsymbol{n}}_i) \right\|_2^2, \tag{2}$$

where N_s denotes the total number of all possible pairs of noisy similar images. First of all, we present the following theorem to justify our Noise2Sim approach:

Noise2Sim Theorem.—Given three sets of independent observations $\{\boldsymbol{s}_i+\boldsymbol{n}_i\}_{i=1}^{N_s}$, $\{\hat{\boldsymbol{n}}_i\}_{i=1}^{N_s}$ *and* $\{\boldsymbol{\delta}_i\}_{i=1}^{N_s}$, *i . e .*, $\forall u, v, u \neq v$, each of these variable pairs $(\boldsymbol{s}_u+\boldsymbol{n}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$, $(\hat{\boldsymbol{n}}_u, \hat{\boldsymbol{n}}_v)$, $(\boldsymbol{\delta}_u, \boldsymbol{\delta}_v)$, $(\hat{\boldsymbol{n}}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$, *and* $(\boldsymbol{\delta}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$ is independent, if the conditional expectation $\forall i$, $\mathbb{E}[\hat{\boldsymbol{n}}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]=0$ and $\mathbb{E}[\boldsymbol{\delta}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]=0$, then the parameters $\boldsymbol{\theta}_s$ optimized in Eq. (2) are asymptotically equivalent (with respect to the limit of the number of training samples) to $\boldsymbol{\theta}_c$ in terms of minimizing the supervised loss in Eq. (1).

Proof.: Let $\boldsymbol{y}_i := f(\boldsymbol{s}_i+\boldsymbol{n}_i;\boldsymbol{\theta})$, expanding the loss function in Eq. (2) and removing the terms that do not involve $\boldsymbol{\theta}$, we have

$$\begin{aligned}\boldsymbol{\theta}_s &= \underset{\theta}{\operatorname{argmin}} \frac{1}{N_s}\sum_{i=1}^{N_s}\left\|\boldsymbol{y}_i-(\boldsymbol{s}_i+\hat{\boldsymbol{n}}_i+\boldsymbol{\delta}_i)\right\|_2^2 \\ &= \underset{\theta}{\operatorname{argmin}} \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\left\|\boldsymbol{y}_i-\boldsymbol{s}_i\right\|_2^2-2\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i-2\boldsymbol{\delta}_i^T\boldsymbol{y}_i\right).\end{aligned} \tag{3}$$

Since $\forall u, v, u \neq v$, each of these variable pairs $(\boldsymbol{s}_u+\boldsymbol{n}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$, $(\hat{\boldsymbol{n}}_u, \hat{\boldsymbol{n}}_v)$, $(\boldsymbol{\delta}_u, \boldsymbol{\delta}_v)$, $(\hat{\boldsymbol{n}}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$, and $(\boldsymbol{\delta}_u, \boldsymbol{s}_v+\boldsymbol{n}_v)$ is independent, then $\forall u, v, u \neq v$, each of these two variable pairs $(\hat{\boldsymbol{n}}_u^T(\boldsymbol{s}_u+\boldsymbol{n}_u), \hat{\boldsymbol{n}}_v^T(\boldsymbol{s}_v+\boldsymbol{n}_v))$ and $(\boldsymbol{\delta}_u T(\boldsymbol{s}_u+\boldsymbol{n}_u), \boldsymbol{\delta}_v T(\boldsymbol{s}_v+\boldsymbol{n}_v))$ is independent; please see Appendix A1 for details. As $\boldsymbol{y}_i = f(\boldsymbol{s}_i+\boldsymbol{n}_i;\boldsymbol{\theta})$ and $f(\cdot;\boldsymbol{\theta})$ is deterministic, then any two variables in each of these two sets $\{\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i\}_{i=1}^{N_s}$ and $\{\boldsymbol{\delta}_i^T\boldsymbol{y}_i\}_{i=1}^{N_s}$ are independent. Since the CT numbers or HU values for patients are within a universal finite interval, usually between [−1000, 3000], any distribution of CT numbers will have finite mean and variance, i.e., the variances $\forall i$, $Var(\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i)$ and $Var(\boldsymbol{\delta}_i^T\boldsymbol{y}_i)$ are finite, According to the weak law of large numbers [52] we have

$$\begin{aligned}\lim_{N_s\to\infty}\frac{1}{N_s}\sum_{i=1}^{N_s}(\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i-\mathbb{E}[\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i]) &\xrightarrow{P} 0 \\ \lim_{N_s\to\infty}\frac{1}{N_s}\sum_{i=1}^{N_s}\left(\boldsymbol{\delta}_i^T\boldsymbol{y}_i-\mathbb{E}\left[\boldsymbol{\delta}_i^T\boldsymbol{y}_i\right]\right) &\xrightarrow{P} 0,\end{aligned} \tag{4}$$

where $\xrightarrow{P}$ means converging in probability. Given $\mathbb{E}[\hat{\boldsymbol{n}}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]=0$ and $\mathbb{E}[\boldsymbol{\delta}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]=0$, we have

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{n}}_i^T\boldsymbol{y}_i] &= \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{n}}_i \mid \boldsymbol{y}_i]^T\boldsymbol{y}_i] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{n}}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]^T\boldsymbol{y}_i] = 0 \\ \mathbb{E}[\boldsymbol{\delta}_i^T\boldsymbol{y}_i] &= \mathbb{E}[\mathbb{E}[\boldsymbol{\delta}_i \mid \boldsymbol{y}_i]^T\boldsymbol{y}_i] = \mathbb{E}[\mathbb{E}[\boldsymbol{\delta}_i \mid \boldsymbol{s}_i+\boldsymbol{n}_i]^T\boldsymbol{y}_i] = 0,\end{aligned} \tag{5}$$

where the first equivalences for both equations in (5) are detailed in Appendix A1, and the second equivalences for both equations in (5) are due to the fact that $\boldsymbol{y}_i = f(\boldsymbol{s}_i+\boldsymbol{n}_i;\boldsymbol{\theta})$ is deterministic. Hence, the following equations hold true:

$$\lim_{N_s \to \infty} \frac{1}{N_s} \sum_{i=1}^{N_s} 2\hat{\boldsymbol{n}}_i^T \boldsymbol{y}_i \xrightarrow{P} 0 \quad \lim_{N_s \to \infty} \frac{1}{N_s} \sum_{i=1}^{N_s} 2\boldsymbol{\delta}_i^T \boldsymbol{y}_i \xrightarrow{P} 0. \tag{6}$$

Given Eqs. (3) and (6) and as $N_s = N_c \to \infty$, we have

$$\begin{aligned} &\operatorname{argmin}_{\theta} \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| \boldsymbol{y}_i - (\boldsymbol{s}_i + \hat{\boldsymbol{n}}_i + \boldsymbol{\delta}_i) \right\|_2^2 \\ &= \operatorname{argmin}_{\theta} \frac{1}{N_c} \sum_{i=1}^{N_c} \left\| \boldsymbol{y}_i - \boldsymbol{s}_i \right\|_2^2. \end{aligned} \tag{7}$$

That is, θ_s and θ_c are asymptotically equivalent (with respect to the limit of the number of training samples) in terms of minimizing the supervised loss function defined in Eq. (1). Note that it is assumed here that the same optimization process is used with the same random seed. □

Noise2Sim vs current deep denoising methods.—Given two practical conditions, zero-mean conditional noise (ZCN), *i.e.*, $\mathbb{E}[\hat{\boldsymbol{n}} \mid \boldsymbol{s} + \boldsymbol{n}] = 0$, and zero-mean conditional discrepancy (ZCD), *i.e.*, $\mathbb{E}[\boldsymbol{\delta}_i \mid \boldsymbol{s}_i + \boldsymbol{n}_i] = 0$, the self-supervised Noise2Sim learning can be regarded as a surrogate of the supervised learning without collecting paired training samples, promising a wide range of applications. During training, Noise2Sim searches and constructs similar sub-images in a non-local manner, while the existing self-supervised methods [37], [38] construct training samples using local pixels only. Noise2Sim enjoys superiority over these self-supervised deep denoising methods as their training samples can be regarded as a subset of those for Noise2Sim, as analyzed in the Appendix A2.

Noise2Sim vs traditional non-local mean methods.—The traditional non-local mean (NLM) method independently processes each pixel in each image via a weighted average of non-local similar pixels. In other words, only its own similar pixels contribute to suppressing noise at the reference pixel in a linear manner. Noise2Sim is different from NLM in two aspects. First, Noise2Sim optimizes a non-linear neural network function to suppress image noises nonlinearly, which is more powerful than linear averaging. Second, instead of separately processing each set of similar pixels, Noise2Sim collaboratively leverages all sets of similar pixels in all training images to learn a denoising network, which is then deployed for the denoising task. Thanks to the high capacity of nonlinear neural networks, all sets of similar pixels instead of a single similar set contribute to recovering each specific pixel.

Suppression of both independent and correlated noises.—Since there is no limitation on the types of noises within sub-images in the theorem, Noise2Sim in principle can suppress both independent and correlated noises as effectively as the supervised deep denoising methods in the limiting case. Also, it is worth mentioning that there are no specific assumptions on the noise distribution. Hence, Noise2Sim can be adapted to process

different noise distributions. Next, we mainly focus on correlated noises in CT applications by searching similar sub-images with the ZCN and ZCD conditions well approximated. The suppression of independent noises in 2D images are introduced in Appendix A2.

B. Suppression of Correlated CT Noises.

Zero-mean conditional noise.—One condition for the Noise2Sim theorem is ZCN. It is well known that CT noise in the image domain is spatially structured, which varies with the scanning protocol parameters, reconstruction steps, patient conditions, scanner hardware factors, etc. Geometrically, each detector row mainly contributes to one or several consecutive slices, and different detector rows are subject to independent noise realizations. As a result, there is a much weaker noise correlation between different slices (especially when the slices are well separated) than within the same slice. Thus, the ZCN condition is practical for CT noises, i.e., $\mathbb{E}[\hat{\boldsymbol{n}}_i \mid \boldsymbol{s}_i + \boldsymbol{n}_i] = \mathbb{E}[\hat{\boldsymbol{n}}_i] = \mathbf{0}$, where the expectation of CT noises is reasonably assumed to be zero [53].

Zero-mean conditional discrepancy.—The other condition for the Noise2Sim theorem to be valid is ZCD. Different from the noise component, the distribution of the discrepancy component $\boldsymbol{\delta}_i$ is not clear. Nevertheless, searching for intrinsically similar sub-images leads to vary small values of $\boldsymbol{\delta}_i$, meaning that the ZCD condition can be well approximated. For medical CT imaging, the same organ or tissue in a patient usually has the similar Hounsfield unit (HU) values and structures, meaning that similar sub-images (patches/slices) can be searched and the ZCD condition can be well approximated as demonstrated in Section III-C.

Independent observation.—In CT applications, $\{(\boldsymbol{s}_i + \boldsymbol{n}_i, \boldsymbol{s}_i + \hat{\boldsymbol{n}}_i + \boldsymbol{\delta}_i)\}_{i=1}^{N_s}$ represent a set of paired similar slices from all patient scans, where $\hat{\boldsymbol{n}}_i$ and $\boldsymbol{\delta}_i$ represent the corresponding noise and discrepancy components respectively in the searched similar slices. Although the slices from the same patient could be somehow correlated, the slices that are far away from each other in the same patient and the slices from different patients should be independent. Then, these observations can be split into multiple sub-sets so that each set is independently observed, i.e., each of these pairs $(\boldsymbol{s}_u + \boldsymbol{n}_u, \boldsymbol{s}_v + \boldsymbol{n}_v)$, $(\hat{\boldsymbol{n}}_u, \hat{\boldsymbol{n}}_v)$, $(\boldsymbol{\delta}_u, \boldsymbol{\delta}_v)$ is independent, where u an v are used to index different paired slices. Also, it is reasonable to assume the noise and discrepancy components of one slice are independent to another independent slice, i.e., each of these two pairs $(\hat{\boldsymbol{n}}_u, \boldsymbol{s}_v + \boldsymbol{n}_v)$ and $(\boldsymbol{\delta}_u, \boldsymbol{s}_v + \boldsymbol{n}_v)$ is independent. It can be regarded as that multiple sub-sets of observations are simultaneously used to optimize Eq. (3), where the independence assumption is practical for each sub-set. Thus, the proof for Noise2Sim Theorem still holds true.

Construction of similar training set.—The volumetric CT image consists of two in-plane dimensions and one through-slice dimension, as shown in Fig. 1 (a). Given the i^{th} reference slice as input, a similar slice can be randomly selected from the $(i-k)^{th}$ to $(i+k)^{th}$ slices, where k defines the searching range of similar slices, and used as the target during training. Similarly, a PCCT image tensor is of four dimensions, including three spatial dimensions (in-plane and through slice), and a channel dimension, as shown in Fig. 1 (b). Thus, the i^{th} slice for PCCT is a 3D tensor (two spatial and one spectral), and a similar

tensor is also randomly selected from the $(i-k)^{th}$ to $(i+k)^{th}$ multi-channel slices. However, it cannot be guaranteed that the pixels/vectors at the same location in different neighboring slices are always similar to each other, especially when the structures are longitudinally changed. According to the Noise2Sim theorem, these dissimilar parts will compromise the zero-mean conditional discrepancy condition, and thus should be excluded from training samples. Here we propose to identify dissimilar pixels between similar slices, which are indicated by the dissimilar mask in Fig. 1 and excluded in computing the loss.

Specifically, we denote a pair of similar LDCT or PCCT images as $\boldsymbol{x}_i, \boldsymbol{x}_j \in R^{H\times W\times C}$, where H, W, C denote height, width, and channel of CT images, $C=1$ for LDCT images, and i, j are the slice indices, $j\in[i-k, i+k]$. For each pair of vectors $\boldsymbol{x}_i(u,v,:)$, $\boldsymbol{x}_j\left(u,v,:\right)\in R^C$ at the same spatial location (u,v), we use their associated patches to determine their similarity. The patches of these two vectors are denoted by $\boldsymbol{P}_i(u,v)$, $\boldsymbol{P}_j\left(u,v\right)\in R^{S\times S\times C}$ that are centered at (u,v) with the patch size $S\times S$. We define the distance map $\boldsymbol{d}\in R^{H\times W}$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as

$$\boldsymbol{d}_{ij}(u,v)=\parallel \boldsymbol{P}_i(u,v)-\boldsymbol{P}_j(u,v)\parallel_2. \tag{8}$$

where $\parallel\cdot\parallel_2$ denotes $L2$ norm. Then, the dissimilar mask $\boldsymbol{m}_{ij}$ is computed as

$$\boldsymbol{m}_{ij}(u,v)=\begin{cases}1 & \boldsymbol{d}_{ij}(u,v)>d_{th}\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

where d_{th} is a predefined threshold.

Loss function.—Finally, the loss function is

$$L=\frac{1}{N_s}\sum_{i,j}\left\|(f(\boldsymbol{x}_i;\boldsymbol{\theta})-\boldsymbol{x}_j)\odot\boldsymbol{m}_{ij}\right\|_2^2, \tag{10}$$

where $\odot$ denotes the element-wise multiplication. In the Noise2Sim theorem, the equivalency between Noise2Sim and Noise2Clean is demonstrated under the MSE loss function. The MSE loss function will drive the network to output the mean value of the noise distribution so that two zero-mean conditions are required. We can also use the L1 loss function to optimize the network, which tends to predict the median value of the noise distribution. If the median of the noise distribution is closer to zero, L1 loss function will achieve better results, as empirically demonstrated in the LDCT denoising experiments.

III. Experiments and Results

A. Datasets

1) Low-dose CT Images: In the main text, we used the publicly available Mayo dataset for the Low Dose CT (LDCT) Grand Challenge[1]. We used eight patients data for training

[1] https://www.aapm.org/grandchallenge/lowdosect/

and the other two for testing. Specifically, the training dataset consists of 4800 512 × 512 slices, and the testing dataset consists of 1136 512 × 512 slices. To test the generalizability of different methods, we used two real scans of an anthropomorphic phantom from FDA[2]. Specifically, we used five volumetric data including 1) a low-dose (25mA) volume with *b40f* kernel and a normal-dose (200mA) volume with *b40f* kernel, 2) a low-dose (25mA) volume with *b60f* kernel and two normal-dose (200mA) volumes with *b40f* kernel, where these two normal-dose volumes contain the same contents but different noise realizations. Each of these volumes contain 408 slices with the image size of 512 × 512.

2) Photon-counting Micro-CT Images: For evaluating the performance of the proposed method on PCCT data, we have scanned a chicken leg phantom and a live mouse animal model for quantitative and qualitative experiments, respectively. The scans are performed on a commercial photon-counting micro-CT scanner (MARS, MARS Bioimaging Ltd., Christchurch, New Zealand) with a cone beam circular scanning geometry. The source is operated at 80kVp/50μA with 1.96mm Al filtration. There are 3 photon-counting detector (PCD) chips stitching side by side each with 128 × 128 pixels and 5 useful thresholds under the charge summing mode forming 5 effective energy bins, and the pixel size is $0.11 \times 0.11 mm^2$. Two lateral translations have been performed during the scan to cover the $40 mm$-diameter field of view, and 1440 views are collected for one rotation at each translation. The 5 energy bins are 7–20, 20–30, 30–47, 47–73 and >73 in keV. Due to the not negligible amount of bad pixels, iterative method is used for reconstruction. Specifically, the reconstructions are performed with 0.1mm isotropic voxels and with simultaneous algebraic reconstruction (SART) algorithm for 200 iterations. The resultant volume is of size 550 × 550 × 130 for each energy bin, and 5 bins in total.

For the quantitative experiment, a chicken leg with bones is placed in a plastic tube for consecutive scans of normal dose (300ms exposure for each view) and low dose (100ms exposure for each view). For the qualitative experiment, a live mouse bearing a tumor model and injected with blood pool contrast agent (ExiTron nano 12000, NanoPET, Berlin, Germany) is scanned with 300ms exposure for each view under anesthesia with isoflurane. Note that iterative reconstruction technique has inherent denoising effect, and usually the effect is strong at a small iteration number and gradually diminishes as the iteration number increases. On the other hand, the resolution gets enhanced (the image becomes sharper) together with the noise along with the increase of iteration. Hence, for merely SART reconstruction 200 iterations are empirically used to obtain a good balance between noise and resolution (references in Fig. 6). For better performance with hybrid use of SART and Noise2sim, the PCCT images are first reconstructed with 600 iterations (inputs in Figs. 6 and 7) to best reserve the resolution and then go through the network to suppress the increased noise to obtain the final clean high-resolution images.

In addition, we scanned a dead mouse on another MARS spectral CT system, which includes a micro x-ray source and a flat-panel PCD as well. This flat-panel PCD is larger and has 660×124 pixels. The emitting x-ray spectrum at 120kVp is divided into the five energy bins: [7.0 32.0], [32.1 43], [43.1 54], [54.1 70] and [70 120]. The distances between

[2] https://wiki.cancerimagingarchive.net/display/Public/Phantom+FDA

the source to the PCD and object are 310 mm and 210 mm, respectively. In this study, we collected 5,760 views with a translation distance 88.32mm for helical scanning. The number of views for one circle is 720. The voxel size in each energy channel was set to $0.1 \times 0.1 \times 0.1 mm^3$. All reconstructed images were performed using the filtered back-projection method. The reconstruction volume is of $667 \times 394 \times 394 \times 5$, i.e., the slice size is 394×394, 667 slices in total, and 5 energy channels in use.

All photon-counting datasets scanned in this study are publicly available on our project page http://chuangniu.info/projects/noise2im/.

B. Implementation Details

In all our experiments, we used a UNet with the depth of 2 [54] with a residual connection as the denoising network in Noise2Sim, which is the same as that used in [37]. In RED-CNN [14] and MAP-NN [15], specified network architectures are designed for LDCT denoising. We used the same data augmentation strategy as in [37], including random cropping followed by random 90°-rotation and mirroring. The Adam [55] algorithm was coupled with the cosine learning rate schedule [56], with the initial learning rate 0.0005. We empirically set the patch size $s = 7$ and the threshold $d_{th} = 30$ in HU. Our method was implemented on the PyTorch[3] deep learning platform. The codes have been made available at https://github.com/niuchuangnn/noise2sim.

C. Empirical estimation of ZCN and ZCD conditions

In Section II, we have analyzed the feasibility of the ZCN and ZCD conditions assumed by the Noise2Sim theorem. Here we aim to practically estimate the values of different loss terms, i.e., $\frac{1}{N_s}\sum_{i=1}^{N_s} \| \boldsymbol{y}_i - \boldsymbol{s}_i \|_2^2 - 2\hat{\boldsymbol{n}}_i^T \boldsymbol{y}_i - 2\boldsymbol{\delta}_i^T \boldsymbol{y}_i$ in Eq. (3) on the Mayo LDCT dataset, where the first term represents the supervised loss, the second and third terms respectively represent the practical values of the ZCN and ZCD conditions. Specifically, $\boldsymbol{y}_i$ is the denoised image predicted by the Noise2Sim model, which was pre-trained with the original Noise2Sim loss using noisy LDCT images only, $\boldsymbol{s}_i$ is the normal-dose CT image, $\hat{\boldsymbol{n}}_i = \hat{\boldsymbol{x}}_i - \hat{\boldsymbol{s}}_i$ denotes the noise component of the similar LDCT image $\hat{\boldsymbol{x}}_i$, $\hat{\boldsymbol{s}}_i$ is the similar NDCT image, and $\boldsymbol{\delta}_i = \boldsymbol{s}_i - \hat{\boldsymbol{s}}_i$ is the difference image between two similar NDCT images. The values of the loss terms over different training epochs are drawn in the Fig. 2. It can be seen that the practical ZCN and ZCD conditions have a very tiny effects on the total loss relative to the supervised loss, and the their values are basically consistent during training. These results demonstrate that the ZCN and ZCD conditions can be piratically well approximated, and minimizing the Noise2Sim loss is essentially minimizing the supervised loss.

D. Noise2Sim for LDCT denoising

We first evaluated Noise2Sim on the commonly used Mayo clinical dataset containing LDCT and NDCT image pairs, demonstrating that Noise2Sim without using any annotation is comparable to and sometimes even better than supervised learning methods. Popular supervised deep LDCT denoising methods RED-CNN [14] and MAP-NN [15] were selected

[3] https://pytorch.org/

for comparison. Noise2Clean that is the plain supervised version of Noise2Sim serves as the baseline. CycleGAN model was selected as the baseline unpaired learning model. Also, we selected BM3D [11] and Nosie2Void [37] for comparison, which are the most popular traditional denoising method and the representative unsupervised deep denoising method respectively. Note that BM3D has the best performance among the traditional denoising methods as demonstrated in [14] for LDCT denoising. Both visual results and quantitative Structural Similarity Index (SSIM) and Peak signal-to-noise ratio (PSNR) results were used to evaluate the denoising performance.

The quantitative results in Table I show that Noise2Sim achieves PSNR measures comparable to and better SSIM values than the supervised learning methods. Indeed, our paired t-Test indicates that the p values are less than 0.001 for the zero average difference hypothesis; that is, the mean results from different methods are significantly different. Therefore, the mean results can be reliably used to compare different methods. The above results suggest that even if the paired training samples are not available, deep neural networks can sill be optimized and applied to improve LDCT quality up to the state-of-the-art supervised performance. Additionally, we evaluated the effects of three factors on Noise2Sim, including dissimilar pixels, loss functions, and training sets, which are detailed in Section II. The default learning setting, denoted by Noise2Sim, identifies and excludes dissimilar pixels in training samples, uses the mean squared error (MSE) loss function, and performs training and testing on separate datasets. In Table I, we tested three variants of Noise2Sim, i.e., Noise2Sim- that includes dissimilar pixels, Noise2Sim-L1 that uses the L1 loss function, and Noise2Sim* that trains the denoising model using the LDCT images in the testing dataset. The denoising performance of Noise2Sim- is evidently worse than Noise2Sim as the zero-mean conditional discrepancy condition required by the Nosie2Sim theorem is violated if the dissimilar pixels are not excluded. On the Mayo dataset, the L1 loss works better than the MSE loss, and the difference between the L1 and L2 losses is analyzed in Section II. Since Noise2Sim requires neither paired noisy nor paired/unpaired clean samples, we also evaluated the Noise2Sim performance for training and testing on a single testing set directly, denoted by Noise2Sim*, with the results summarized in Table I. Noise2Sim* achieved slightly better results than Noise2Sim likely due to that there is no generalizability issue for our training and testing on the same testing dataset. In this setting, we might not need to collect a large number of extra training samples. However, in this case the time required for denoising covers both training and testing stages, thus being much longer than directly applying the pre-trained model. In practice, either training strategy for Noise2Sim can be selected according to specific application requirements.

The visual results in Fig. 3 also illustrate that Noise2Sim is better than the supervised methods in preserving structural details, as indicated by the zoomed ROIs in the yellow bounding boxes, in reference to the normal-dose CT image. Clearly, the supervised methods tend to remove noises aggressively while Noise2Sim tries to preserve informative details. As a weakly-supervised learning model trained on unpaired normal-dose CT images, CycleGAN suppresses CT image noise with the overall appearance of the resultant images similar to the NDCT images. However, CycleGAN tends to shift the HU values and sometimes generates fake structural details as shown in Fig. 3. Our results show that

Noise2Void does not work for structured CT noises, and the BM3D results are either over-smoothed or compromised with structured artifacts. In contrast, Noise2Sim has significantly better performance in reducing structured noises and preserving content structures.

E. Controllability of Noise2Sim with denoising levels

The similarity parameter k defined in Section II allows us to control the extent to which image noise is removed. As shown in Fig 4, more noise can be suppressed with a larger parameter k. A larger k increases the noise independence of similar training samples so that noise can be suppressed more aggressively, as implied by the Noise2Sim theorem. On the other hand, the denoising results may be harmed if k is too large, as the zero-mean conditional discrepancy assumed by the Noise2Sim theorem may be compromised. The statistical results in Fig 4 shows that $k = 2$ achieves the best trade-off between noise independence and feature similarity. The parameter k can be adjusted according to specific downstream tasks. If image quality can be quantitatively modeled, such as with a neural network and/or a Gram matrix [57], k could be automatically optimized. More practically, several levels of denoising images can be simultaneously presented to radiologists, and thus the best image quality can be determined with the human expertise in loop. Note that the denoising level is not controlled on-the-fly using Noise2Sim. In fact, different denoising levels can be achieved by training several denoising models. Since the denoising model is relatively small, multiple models can be trained and deployed in parallel.

F. Effects of Threshold and Patch size

As described in Section II-B, we introduced two hyperparameters including the threshold d_{th} and patch size $\mathcal{S}$ in computing the similarity. Here we empirically evaluated their effects on low-dose CT denoising, and the results are reported in Tables II and III. The results show that the denoising performance was evidently degraded when $d_{th} = 0$; *i.e.*, all dissimilar pixels were included during the training process. Other values (10 ~ 60) close to the default value (30) had a tiny impact on the results. On the other hand, the denoising results for the patch size of 3 × 3 was clearly worse than that for larger values, with 7 × 7 patch size achieving the best results.

G. Generalizability of Noise2Sim on LDCT datasets

We further evaluated Noise2Sim on real LDCT scans of the FDA anthropomorphic phantom [58], demonstrating that Noise2Sim performs better than the supervised learning methods that transfer from the simulated to real data. Since we can hardly obtain paired LDCT and NDCT images in the clinical scenario, we applied the supervised learning models that trained on the Mayo dataset to processing the real LDCT images in the FDA dataset. For this purpose, Noise2Sim was trained and tested using the same LDCT images in the FDA dataset. Specifically, we used two low-dose (25mA) phantom scans that were reconstructed with different kernels denoted by *b40f* and *b60f*, resulting in different noise patterns respectively. To quantify the denoising performance, we used the corresponding normal-dose (200mA) phantom scans as targets. Because the normal-dose images reconstructed with the *b40f* kernel still contain noises, we first processed the normal-dose images by training the Noise2Noise model with the paired noisy images and used the denoised noraml-dose

image as the reference. Table IV shows that Noise2Sim achieves the best results on both two phantoms in terms of PSNR and SSIM among all the supervised and unsupervised learning methods. Particularly, the supervised learning models cannot work well when being transferred to denoising real LDCT images that are reconstructed with a totally different kernel *b40f*. The *b40f* phantom images denoised using different methods is visualized in Fig. 5. These results show that supervised learning models suffer from severe generalizability issues when there is a shift in distributions between training and testing samples. In contrast, Noise2Sim can be directly optimized in the target domain without suffering from the generalizability issue. Surprisingly, Noise2Sim can recover underlying structures despite strong noises, while these structures can be hardly seen in the original LDCT images. These strongly demonstrate the effectiveness of the proposed unsupervised deep denoising approach, which has a great potential to improve the LDCT image quality and diagnostic performance.

H. Noise2Sim for PCCT denoising

As an emerging CT imaging technology, an x-ray photon-counting detector records individual x-ray photons as well as their energy levels. By the nature of the photon-counting mechanism, this technique is free of electronic noise and provides a much higher resolution compared to that of conventional energy integrating detectors [59]. The x-ray energy information can be used to extract chemically specific information and facilitate beam hardening correction, metal artifact reduction, tissue characterization and K-edge imaging [60], [61]. On the other hand, the refinement of spatial resolution and the division of photons into narrow energy bins significantly raise the image noise in individual energy channels, demanding powerful denoising methods that remove image noise while preserving fine features.

To evaluate the effectiveness of Noise2Sim on denoising PCCT images, we scanned a chicken drumstick at low- and normal-dose setting respectively and reconstructed them into PCCT image volumes using the state-of-the-art MARS spectral CT system. Although the low-dose and normal-dose PCCT images of the same object are available, it is still hard to directly apply supervised learning methods for PCCT image denoising since different scans were not in perfect registration. Thus, we first used a registration tool to align the low-dose and normal-dose PCCT volumes so that the supervised leaning method (Noise2Clean) can be applied. Then, the quantitative results were calculated for comparison. For this purpose, we still selected the unsupervised BM3D method as the baseline. Fig. 6 shows that our proposed unsupervised leaning method outperforms BM3D and even the supervised learning method in terms of both the quantitative and qualitative results. BM3D requires manually tuning the prior parameter of standard deviation (std) for each bin to achieve decent results. Despite these tedious tuning steps, BM3D still produced either over-smoothed features (the 4^{st} bin with the std of 0.2) or structured artifacts (the 5^{th} bin with std of 0.5). The reason seems that BM3D could either blindly remove high-frequency components including some fine structures in the transform domain according to the patch similarity to address a large standard deviation prior or preserve some large structured noises to reflect a small standard deviation prior. On the other hand, due to the imperfect alignment of low-dose and normal-dose training samples, the performance of a supervised learning method may

be degraded. In contrast, Noise2Sim intrinsically registers similar training samples and can be then applied to produce excellent denoising results without tuning the hyperparameter in a bin-specific fashion. Particularly, Noise2Sim is able to faithfully recover the underlying structures interfered by strong noises, as shown in the last row of Fig. 6, which is visually even better than the reference image.

Finally, we applied Noise2Sim to a PCCT image volume of a live mouse scanned on the MARS spectral CT system. In this case, only a single normal-dose scan was available so that the supervised leaning method cannot be applied. Fig. 7 shows this challenging case in the 5^{th} bin, where structures can be hardly observed in the original noisy image. BM3D can only recover a rough contour of the spine. In contrast, Noise2Sim can accurately recover the spine and other organs. Also, more comparison results can be found in Fig. 8, showing that the proposed method is consistently better than BM3D in terms of different denoising levels on PCCT images.

IV. Discussion and Conclusion

In the Noise2Sim method, the key is to find similarity-based training samples satisfying the ZCN and ZCD conditions from noisy data without any labels. Specific to the dimensionality of involved images, several efficient searching algorithms have been evaluated for suppressing noises in this study. Since here we focus on unsupervised learning, the commonly used UNet [54] architecture and the Euclidean distance were applied for similarity measurement. Clearly, the denoising performance should be better with a more advanced network design, more accurate similarity measurement, and a more sophisticated loss function.

As a general unsupervised denoising approach, Noise2Sim can be adapted to many other domains, not limited to CT images. In the general spirit of Noise2Sim, deeper analysis of domain-specific data can help fully leverage similar data features and achieve superior performance. As far as LDCT denoising is concerned, the imaging performance may be improved using a dual domain denoising network with similarity matches in both the sinogram and image domains synergistically. Moreover, due to its simplicity and efficiency, our proposed approach could be also incorporated into the classic model-based reconstruction algorithms [50], [62], [63] as an element or a constraint to regularize the CT image reconstruction process.

In conclusion, we have presented a novel similarity-based self-supervised learning denoising approach only using noisy images with neither clean nor noisy paired labels. To our best knowledge, our proposed denoising approach is the first-of-its-kind to suppress structured noises in the self-supervised deep learning fashion. Theoretically, we have proved the Noise2Sim theorem that the similarity-based self-supervised learning method is asymptotically equivalent to the supervised learning methods under mild conditions. Also, we have applied the Noise2Sim approach in the important applications to denoise 3D low-dose CT and 4D photon-counting micro-CT images. Our experimental results have demonstrated the superiority of the Noise2Sim approach in comparison with existing model-

based and deep learning denoising methods. Potentially, the Noise2Sim approach can be adapted to various domains for practical denoising performance.

Acknowledgments

This work was supported in part by NIH/NCI under Award numbers R01CA233888, R01CA237267, R21CA264772, and NIH/NIBIB under Award numbers R01EB026646, R01HL151561, R01EB031102.

Appendix

A1. Theoretical Analysis

Independence.

Given four variables X_1, X_2, Y_1, Y_2, if each of these pairs (X_1, X_2), (Y_1, Y_2), (X_1, Y_2), (Y_1, X_2) is independent, then we can we have

$$\begin{aligned} &P(X_1, X_2, Y_1, Y_2) \\ &= P(X_1 \mid X_2, Y_1, Y_2)P(X_2 \mid Y_1, Y_2)P(Y_1 \mid Y_2)P(Y_2) \\ &= P(X_1 \mid Y_1)P(X_2 \mid Y_2)P(Y_1)P(Y_2) \\ &= P(X_1, Y_1)P(X_2, Y_2) \end{aligned} \tag{A1}$$

Thus, X_1Y_1 and X_2Y_2 are independent. Corresponding to the theorem proof in Section II, let $X_1 := \boldsymbol{s}_u + \boldsymbol{n}_u$, $X_2 := \boldsymbol{s}_v + \boldsymbol{n}_v$, $Y_1 = \hat{\boldsymbol{n}}_u$, $Y_2 = \hat{\boldsymbol{n}}_v$, then the variables $\hat{\boldsymbol{n}}_u^T(\boldsymbol{s}_u + \boldsymbol{n}_u)$ and $\hat{\boldsymbol{n}}_v^T(\boldsymbol{s}_v + \boldsymbol{n}_v)$ are independent. If we let $Y_1 := \boldsymbol{\delta}_u$ and $Y_2 := \boldsymbol{\delta}_v$, then the variables $\boldsymbol{\delta}_u^T(\boldsymbol{s}_u + \boldsymbol{n}_u)$ and $\delta_v^T(\boldsymbol{s}_v + \boldsymbol{n}_v)$ are independent.

Expectation equivalence.

Given two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, then we have

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{X}^T\boldsymbol{Y}\right] &= \int_x \int_y \boldsymbol{x}^T\boldsymbol{y} f_{XY}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \\ &= \int_x \int_y \boldsymbol{x}^T\boldsymbol{y} f_{X\mid Y}(\boldsymbol{x} \mid \boldsymbol{y}) f_Y(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \\ &= \int_x \int_y \boldsymbol{x}^T\boldsymbol{y} f_{X\mid Y}(\boldsymbol{x} \mid \boldsymbol{y}) f_Y(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \\ &= \int_y \left(\int_x \boldsymbol{x} f_{X\mid Y}(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{x}\right)^T \boldsymbol{y} f_Y(\boldsymbol{y}) d\boldsymbol{y} \\ &= \mathbb{E}[\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}]^T\boldsymbol{Y}] \end{aligned} \tag{A2}$$

Corresponding to Eqs. (5), if we let $X := \hat{\boldsymbol{n}}_i$ or $X := \boldsymbol{\delta}_i$ and $Y := \boldsymbol{y}_i$, then the first equivalences hold true.

A2. Suppression of independent noises in 2D images

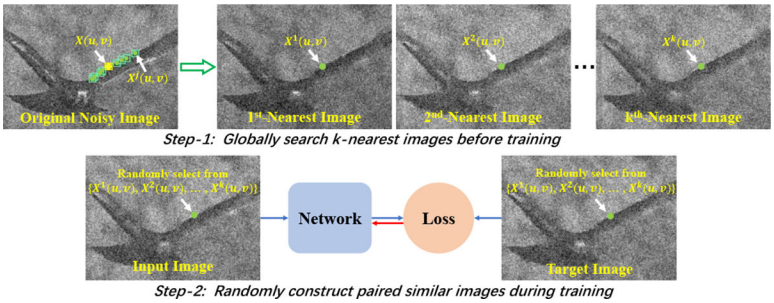

Fig. A1.

Noise2Sim training process on 2D images with independent noise. Step 1 is to search for a set of k similar pixels for each pixel in the original noisy image, and form k most similar images, which is also referred to as nearest images. In the original noisy image, the yellow and green points respectively denote the reference pixel $\boldsymbol{X}(u, v)$ and its k nearest pixels $\boldsymbol{X}^j(u, v)$, $j = 1, \cdots, k$, and the corresponding boxes present the patches respectively centered at these involved pixels that are used to compute the similarity between center pixels. The k^{th} nearest image is formed by replacing all pixels in the original noisy image with their k^{th} nearest pixels. Step 2 is to construct a pair of similar images as the input and the target to train a deep neural network. Each similar image is independently constructed by replacing every pixel $\boldsymbol{X}(u, v)$ in the original noisy image with the pixel randomly selected from $\mathcal{N}(u, v) = \{\boldsymbol{X}(u, v), \boldsymbol{X}^1(u, v), \cdots, \boldsymbol{X}^k(u, v)\}$, which are pre-computed in Step 1.

In the case of 2D images corrupted by independent noise, we design an efficient two-step algorithm to construct a large set of similar images from noisy images only. As introduced in Section II-A, the similar training set is defined as $\{(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_i) \mid S(T(\boldsymbol{x}_i), T(\hat{\boldsymbol{x}}_i))\}$, where T is a transformation function for sub-images, and S is a function to measure between two sub-images. Here the similar sub-images are constructed by replacing each and every original pixels with the searched similar pixels, where the similarity between two pixels are measured by their surrounding patches. For simplicity, here T is the identity function that means no transformations are applied to image patches, and one can also use other transformations to reduce the variance of similarity estimation caused by noises [11]. For the similarity estimation S, we adopt the k – NN strategy that each pixel is matched with k nearest similar pixels in terms of the Euclidean distance between their surrounding patches.

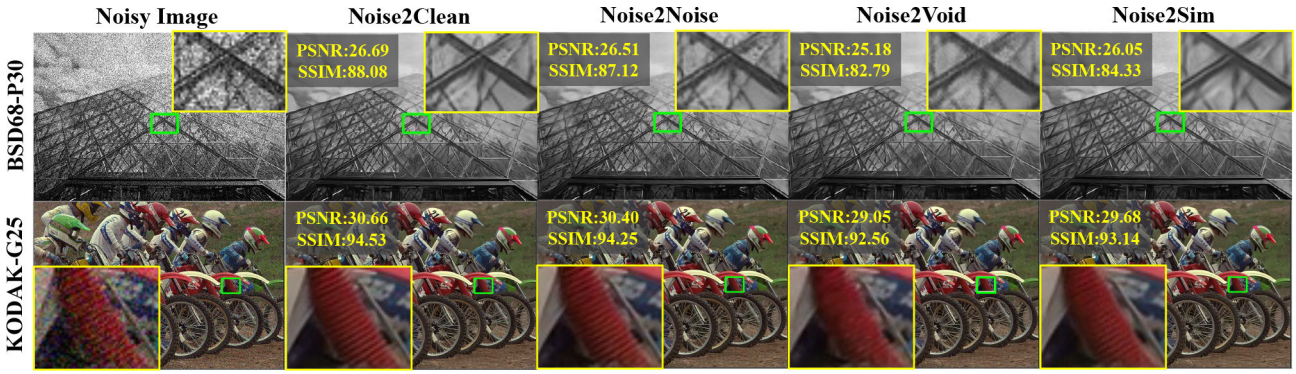

Fig. A2.
Visual comparison of denoising results on different datasets, where $G25$ means Gaussian noise with $\boldsymbol{Std} = \mathbf{25}$, and $P30$ means Poisson noise with $\lambda = 30$. The PSNR and SSIM values are included.

Specifically, let us describe this process in detail at the pixel-level. For each reference pixel $\boldsymbol{x}(u, v)$ with its coordinates (u, v) in a given noisy image $\boldsymbol{x}$, we compute its k nearest pixels over the whole image. The distance between two pixels $\boldsymbol{x}(u_1, v_1)$ and $\boldsymbol{x}(u_2, v_2)$ is defined as the Euclidean distance between their associated patches; i.e., $\| \boldsymbol{S}(u_1, v_1) - \boldsymbol{S}(u_2, v_2) \|_2$, where $\boldsymbol{S}(u, v)$ denotes a square patch defined by the pre-defined patch size and the center pixel $\boldsymbol{x}(u, v)$. Thus, each position in the image has a set of $k + 1$ similar pixels (+1 means the reference pixel included), denoted as $\mathcal{N}(u, v) = \{\boldsymbol{x}(u, v), \boldsymbol{x}^1(u, v), \cdots, \boldsymbol{x}^k(u, v)\}$, where $\boldsymbol{x}^j(u, v)$ denotes the j-th nearest pixel relative to $\boldsymbol{x}(u, v)$. As shown in Fig. A1, the yellow and green dots denote the reference pixel and its nearest pixels, and the boxes are the associated patches. Based on these similar pixel sets, a similar noisy image can be constructed by replacing every original pixel $\boldsymbol{x}(u, v)$ with a similar one randomly selected from $\mathcal{N}(u, v)$. Then, a pair of similar images are independently constructed in each iteration.

Noise2Sim can be regarded as a non-local version of Noise2Void leveraging both local and global similarity information, while Noise2Void only uses neighbor pixels to construct a target. Therefore, Noise2Sim enjoys principled superiority over Noise2Void.

Furthermore, Noise2Sim was evaluated on the commonly used BSD68 and KODAK images corrupted by Gaussian and Poisson noises. All implementations on 2D images can be found at http://chuangniu.info/projects/noise2im/. Consistent to the known limitations discussed in [37], some grainy structures cannot be recovered well in Noise2Void denoised images, as shown in the zoom-in regions of the last row in Fig. A2. In contrast, our Noise2Sim images well preserve structural subtleties, as self-similarities are fully utilized in the training stages.

References

[1]. Hsieh SS, Leng S, Rajendran K, Tao S, and McCollough CH, “Photon counting CT: Clinical applications and future developments,” IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 5, no. 4, pp. 441–452, 2021. [PubMed: 34485784]

[2]. Berrington de González A, Mahesh M, Kim K-P, Bhargavan M, Lewis R, Mettler F, and Land C, “Projected Cancer Risks From Computed Tomographic Scans Performed in the United States

in 2007," Archives of Internal Medicine, vol. 169, no. 22, pp. 2071–2077, 12 2009. [PubMed: 20008689]

[3]. Lindenbaum M, Fischer M, and Bruckstein A, "On Gabor's contribution to image enhancement," Pattern Recognition, vol. 27, no. 1, pp. 1 – 8, 1994.

[4]. Perona P and Malik J, "Scale-space and edge detection using anisotropic diffusion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 629–639, 1990.

[5]. Catté F, Lions P-L, Morel J-M, and Coll T, "Image selective smoothing and edge detection by nonlinear diffusion," SIAM Journal on Numerical Analysis, vol. 29, no. 1, pp. 182–193, 1992.

[6]. Yaroslavsky LP, "Digital picture processing - an introduction." Springer Verlag, 1985.

[7]. Smith SM and Brady JM, "SUSAN-A New Approach to Low Level Image Processing," Int. Journal of Computer Vision, vol. 23, no. 1, pp. 45–78, 1997.

[8]. Tomasi C and Manduchi R, "Bilateral filtering for gray and color images," in ICCV, 1998, pp. 839–846.

[9]. Yu G and Sapiro G, "DCT image denoising: a simple and effective image denoising algorithm," Image Processing On Line, vol. 1, pp. 292–296, 2011.

[10]. Buades A, Coll B, and Morel J, "A non-local algorithm for image denoising," in CVPR, vol. 2, 2005, pp. 60–65 vol. 2.

[11]. Dabov K, Foi A, Katkovnik V, and Egiazarian K, "Image denoising by sparse 3-d transform-domain collaborative filtering," IEEE Transactions on Image Processing, vol. 16, no. 8, pp. 2080–2095, 2007. [PubMed: 17688213]

[12]. Zhang K, Zuo W, Chen Y, Meng D, and Zhang L, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142–3155, 2017. [PubMed: 28166495]

[13]. Tian C, Fei L, Zheng W, Xu Y, Zuo W, and Lin C-W, "Deep learning on image denoising: An overview," Neural Networks, vol. 131, pp. 251–275, 2020. [PubMed: 32829002]

[14]. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, and Wang G, "Low-dose CT with a residual encoder-decoder convolutional neural network," IEEE Transactions on Medical Imaging, vol. 36, no. 12, pp. 2524–2535, 2017. [PubMed: 28622671]

[15]. Shan H, Padole A, Homayounieh F, Kruger U, Doda Khera R, Nitiwarangkul C, Kalra MK, and Wang G, "Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction," Nature Machine Intelligence, vol. 1, p. 269–276, 2019.

[16]. Zhang W, Zhou Z, Gao Z, Yang G, Xu L, Wu W, and Zhang H, "Multiple adversarial learning based angiography reconstruction for ultra-low-dose contrast medium ct," IEEE Journal of Biomedical and Health Informatics, 2022.

[17]. Wu W, Guo X, Chen Y, Wang S, and Chen J, "Deep embedding-attention-refinement for sparse-view ct reconstruction," IEEE Transactions on Instrumentation and Measurement, pp. 1–1, 2022.

[18]. Zeng D, Huang J, Bian Z, Niu S, Zhang H, Feng Q, Liang Z, and Ma J, "A simple low-dose x-ray CT simulation from high-dose scan," IEEE transactions on nuclear science, vol. 62, no. 5, pp. 2226–2233, 2015. [PubMed: 26543245]

[19]. Niu C and et al. , "Noise entangled GAN for low-dose CT simulation," arXiv preprint arXiv:2102.09615, 2021.

[20]. Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, and Aila T, "Noise2Noise: Learning image restoration without clean data," in ICML, vol. 80, 2018, pp. 2965–2974.

[21]. Hasan AM, Mohebbian MR, Wahid KA, and Babyn P, "Hybrid-collaborative Noise2Noise denoiser for low-dose CT images," IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 5, no. 2, pp. 235–244, 2020.

[22]. Won D, Jung E, An S, Chikontwe P, and Park SH, "Self-supervised learning based CT denoising using pseudo-CT image pairs," arXiv preprint arXiv:2104.02326, 2021.

[23]. Wu D, Kim K, and Li Q, "Low-dose CT reconstruction with Noise2Noise network and testing-time fine-tuning," Medical Physics, vol. 48, no. 12, pp. 7657–7672, 2021. [PubMed: 34791655]

[24]. Yuan N, Zhou J, and Qi J, “Half2half: deep neural network based CT image denoising without independent reference data,” Physics in Medicine and Biology, vol. 65, no. 21, p. 215020, 2020. [PubMed: 32707565]

[25]. Zhang C, Chang S, Bai T, and Chen X, “S2ms: Self-supervised learning driven multi-spectral CT image enhancement,” arXiv preprint arXiv:2201.10294, 2022.

[26]. Fang W, Wu D, Kim K, Kalra MK, Singh R, Li L, and Li Q, “Iterative material decomposition for spectral CT using self-supervised Noise2Noise prior,” Physics in Medicine and Biology, vol. 66, no. 15, p. 155013, 2021.

[27]. Zhang Z, Liang X, Zhao W, and Xing L, “Noise2Context: Context-assisted learning 3d thin-layer for low-dose CT,” Medical Physics, 2021.

[28]. Kim B, Shim H, and Baek J, “Weakly-supervised progressive denoising with unpaired CT images,” Medical Image Analysis, vol. 71, p. 102065, 2021.

[29]. Kwon T and Ye JC, “Cycle-free CycleGAN using invertible generator for unsupervised low-dose CT denoising,” IEEE Transactions on Computational Imaging, vol. 7, pp. 1354–1368, 2021.

[30]. Li Z, Zhou S, Huang J, Yu L, and Jin M, “Investigation of low-dose CT image denoising using unpaired deep learning methods,” IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 5, no. 2, pp. 224–234, 2021. [PubMed: 33748562]

[31]. Gu J and Ye JC, “Adain-based tunable CycleGAN for efficient unsupervised low-dose CT denoising,” IEEE Transactions on Computational Imaging, vol. 7, pp. 73–85, 2021.

[32]. Niu C and Wang G, “Unsupervised contrastive learning based transformer for lung nodule detection,” arXiv preprint arXiv:2205.00122, 2022.

[33]. Niu C, Zhang J, Wang G, and Liang J, “GATCluster: Self-supervised Gaussian-attention network for image clustering,” in European Conference on Computer Vision, 2020, pp. 735–751.

[34]. Niu C and Wang G, “HOME: High-order mixed-moment-based embedding for representation learning,” arXiv preprint arXiv:2207.07743, 2022.

[35]. Niu C, Shan H, and Wang G, “SPICE: Semantic pseudo-labeling for image clustering,” arXiv preprint arXiv:2103.09382, 2021.

[36]. Niu C and Wang G, “Self-supervised representation learning with multi-segmental informational coding (MUSIC),” arXiv preprint arXiv:2206.06461, 2022.

[37]. Krull A, Buchholz T, and Jug F, “Noise2Void - learning denoising from single noisy images,” in CVPR, 2019, pp. 2124–2132.

[38]. Batson J and Royer L, “Noise2Self: Blind denoising by self-supervision,” in ICML, vol. 97, 2019, pp. 524–533.

[39]. Xie Y, Wang Z, and Ji S, “Noise2Same: Optimizing a self-supervised bound for image denoising,” in NeurIPS, vol. 33, 2020, pp. 20 320–20 330.

[40]. Krull A, Vičar T, Prakash M, Lalit M, and Jug F, “Probabilistic Noise2Void: Unsupervised content-aware denoising,” Frontiers in Computer Science, vol. 2, p. 5, 2020.

[41]. Khodja A, Zheng Z, and He Y, “Similarity noise training for image denoising,” Mathematics and Computer Science, vol. 5, no. 2, pp. 56—63, 2020.

[42]. Laine S, Karras T, Lehtinen J, and Aila T, “High-quality self-supervised deep image denoising,” in NeurIPS, 2019, pp. 6970–6980.

[43]. Broaddus C, Krull A, Weigert M, Schmidt U, and Myers G, “Removing structured noise with self-supervised blind-spot networks,” in International Symposium on Biomedical Imaging, 2020, pp. 159–163.

[44]. Lempitsky V, Vedaldi A, and Ulyanov D, “Deep image prior,” in CVPR, 2018, pp. 9446–9454.

[45]. Quan Y, Chen M, Pang T, and Ji H, “Self2self with dropout: Learning self-supervised denoising from single image,” in CVPR, June 2020.

[46]. Kim K, Soltanayev S, and Chun SY, “Unsupervised training of denoisers for low-dose CT reconstruction without full-dose ground truth,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 6, pp. 1112–1125, 2020.

[47]. Harte J, Kinzig A, and Green J, “Self-similarity in the distribution and abundance of species,” Science, vol. 284, no. 5412, pp. 334–336, 1999. [PubMed: 10195901]

[48]. Niu C and Wang G, “Noise2Sim – similarity-based self-learning for image denoising,” 2011.03384 v1, 2020.

[49]. Chen J and et al. , “Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes,” Nature Methods, vol. 18, p. 678–687, 2021. [PubMed: 34059829]

[50]. Li Y, Bar-Shira O, Monga V, and Eldar YC, “Deep algorithm unrolling for biomedical imaging,” arXiv preprint arXiv:2108.06637, 2021.

[51]. Fang W, Fu L, and Li H, “Unsupervised CNN based on self-similarity for seismic data denoising,” IEEE Geoscience and Remote Sensing Letters, pp. 1–1, 2021.

[52]. Durrett R, Probability: theory and examples. Cambridge university press, 2019, vol. 49.

[53]. Divel SE and Pelc NJ, “Accurate image domain noise insertion in ct images,” IEEE transactions on medical imaging, vol. 39, no. 6, pp. 1906–1916, 2019. [PubMed: 31870981]

[54]. Ronneberger O, Fischer P, and Brox T, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI N, Navab J, Hornegger W, Wells M, and Frangi AF, Eds., 2015, pp. 234–241.

[55]. Kingma DP and Ba J, “Adam: A method for stochastic optimization,” in ICLR, Bengio Y and LeCun Y, Eds., 2015.

[56]. Loshchilov I and Hutter F, “Sgdr: Stochastic gradient descent with warm restarts,” in ICLR, 2017.

[57]. Sreeram V and Agathoklis P, “On the properties of gram matrix,” IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 41, no. 3, pp. 234–237, 1994.

[58]. Gavrielides MA, Kinnard LM, Myers KJ, Peregoy J, Pritchard WF, Zeng R, Esparza J, Karanian J, and Petrick N, “A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom,” Opt. Express, vol. 18, no. 14, pp. 15 244–15 255, Jul 2010.

[59]. Li M, Fang Z, Cong W, Niu C, Wu W, Uher J, Bennett J, Rubinstein JT, and Wang G, “Clinical micro-CT empowered by interior tomography, robotic scanning, and deep learning,” IEEE Access, vol. 8, pp. 229 018–229 032, 2020.

[60]. Wu W, Hu D, Niu C, Broeke LV, Butler AP, Cao P, Atlas J, Chernoglazov A, Vardhanabhuti V, and Wang G, “Deep learning based spectral CT imaging,” Neural Networks, 2021.

[61]. Wu W, Hu D, An K, Wang S, and Luo F, “A high-quality photon-counting ct technique based on weight adaptive total-variation and image-spectral tensor factorization for small animals imaging,” IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–14, 2021. [PubMed: 33776080]

[62]. Noo F, Clackdoyle R, and Pack JD, “A two-step hilbert transform method for 2d image reconstruction,” Physics in Medicine & Biology, vol. 49, no. 17, p. 3903, 2004. [PubMed: 15470913]

[63]. Noo F, Defrise M, Clackdoyle R, and Kudo H, “Image reconstruction from fan-beam projections on less than a short scan,” Physics in medicine & biology, vol. 47, no. 14, p. 2525, 2002. [PubMed: 12171338]

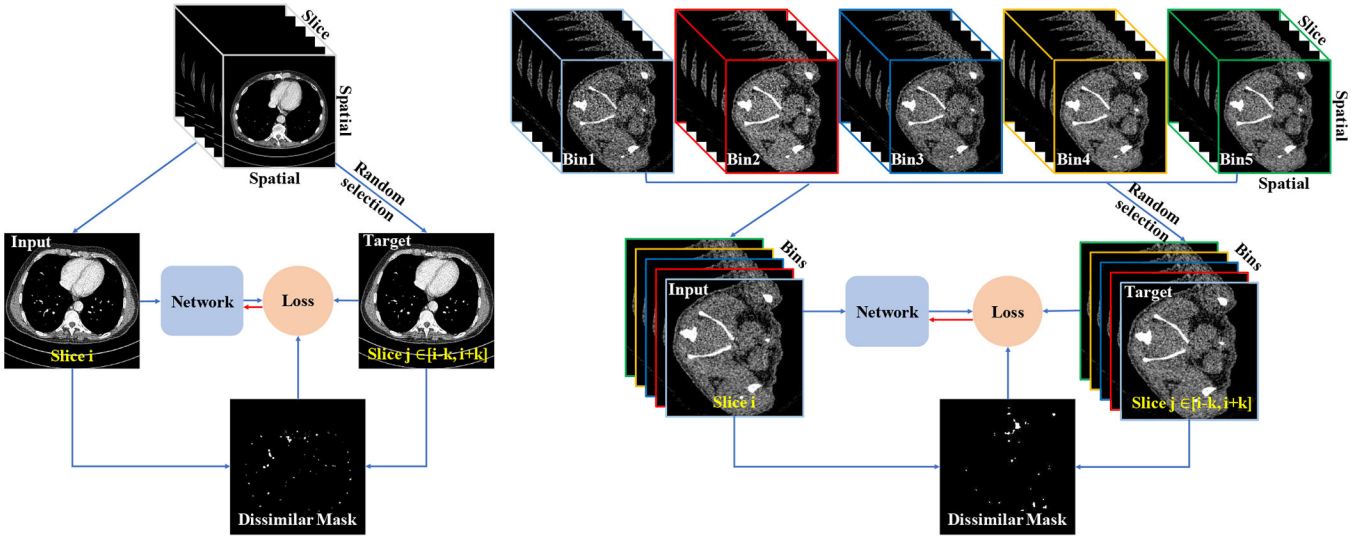

Fig. 1.

Noise2Sim training process on LDCT and PCCT. The similar volumes along the slice direction are selected to construct training samples, where the dissimilar vectors identified in a mask image are excluded during training. For PCCT, different colors denote different bins, so both inputs and targets are multi-channel slices from multiple energy bins.

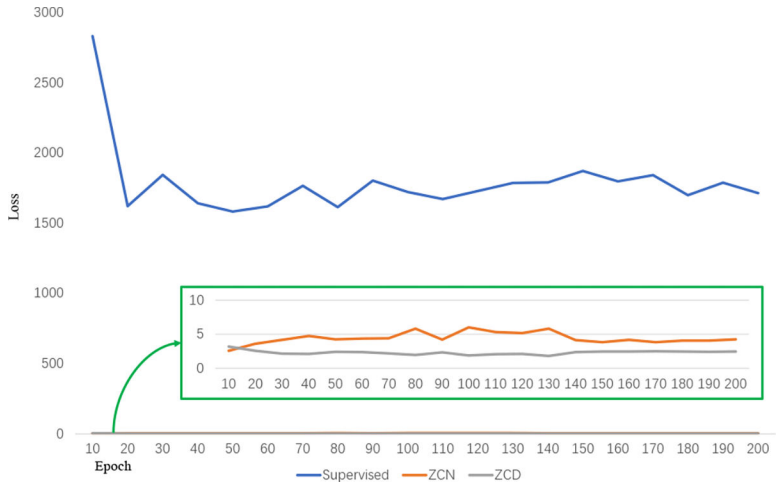

Fig. 2.

Values of different loss terms. The zoomed-in green box shows the detailed values of the ZCN and ZCD terms.

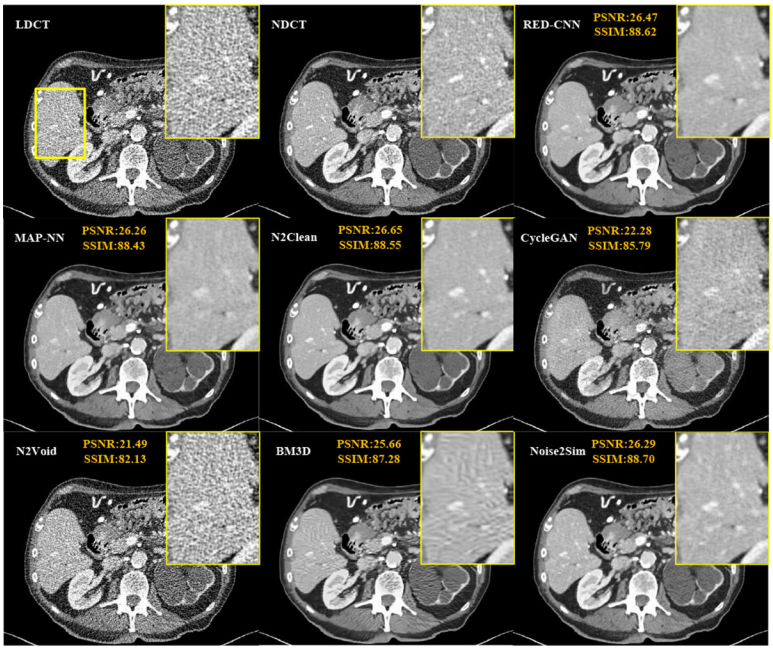

Fig. 3.

Superior/comparative Noise2Sim results on the Mayo LDCT dataset. The yellow ROIs indicate that detail structures are better preserved using Noise2Sim than other methods in reference to the normal-dose CT image. The display window is [−160, 240] in Hounsfield unit (HU), along with PSNR and SSIM values.

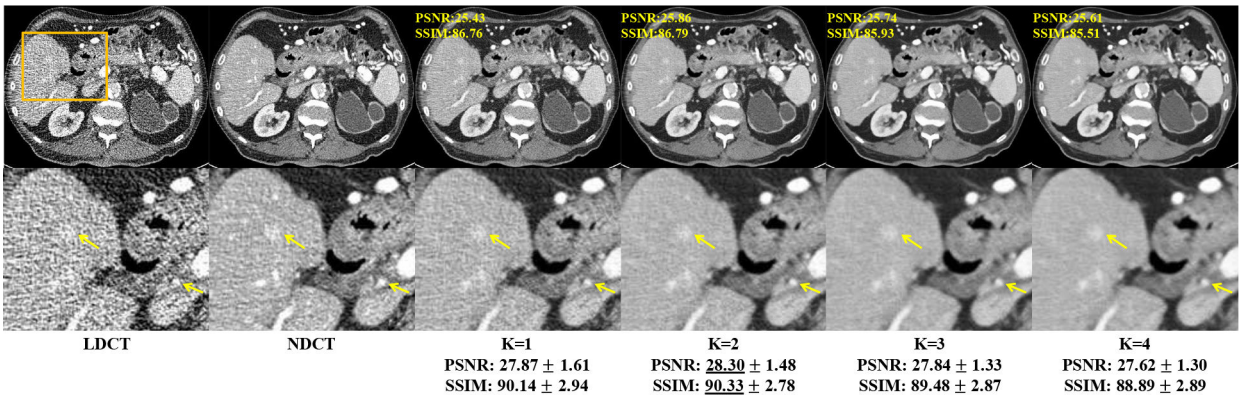

Fig. 4.

Controllable Noise2Sim results with different denoising levels on the Mayo dataset. The numbers in yellow and black are calculated for the individual case. The numbers in black are respectively the mean and standard deviation values over the test dataset (1136 slices). The second row shows the corresponding ROIs in the yellow bounding boxes. The yellow arrows indicate that structural details are enhanced via image denoising.

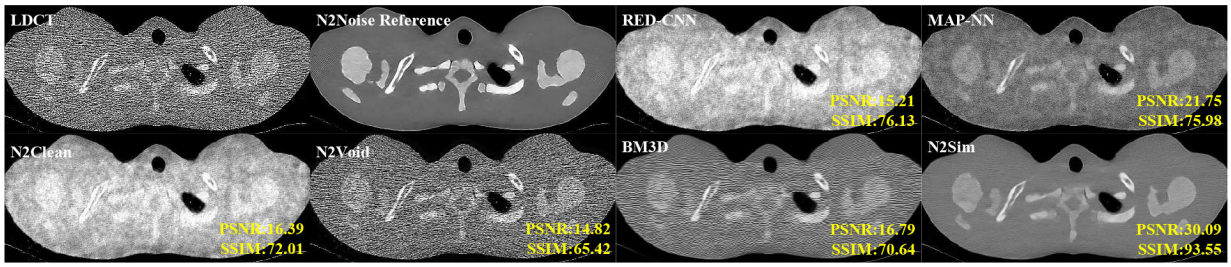

Fig. 5.

Generalizability results of Noise2Sim on the FDA dataset, significantly better than those from other denoising methods. The LDCT image was obtained with 25mA and *b40f* kernel. The N2Noise reference image was obtained with 200mA and processed with the paired Noise2Noise method. The display window is [−160, 240] HU, along with PSNR and SSIM values.

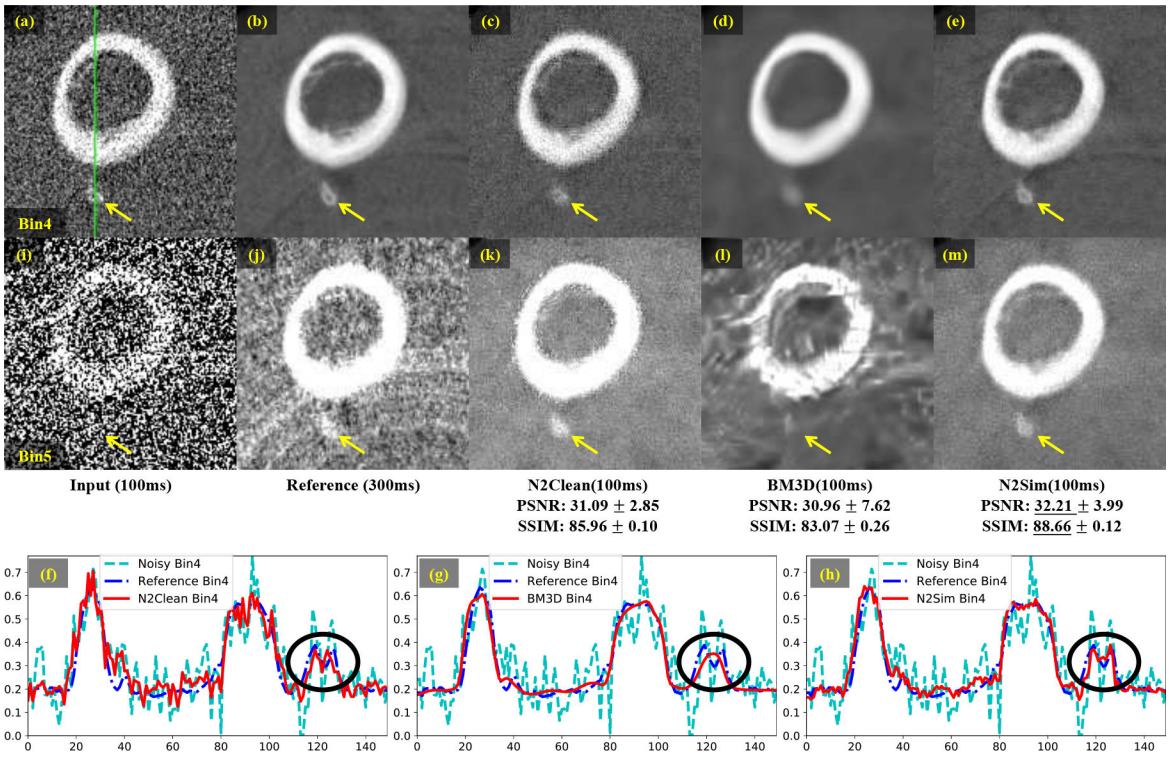

Fig. 6.

Qualitative and quantitative results of Noise2Clean, BM3D, and Noise2Sim on 4D photon-counting CT images. (a)-(e) and (d)-(m) respectively show denoising results from the **4*th*** and **5*th*** energy bins. (f)-(h) show the corresponding profiles along the green line in (a). The PSNR and SSIM results over the test dataset are reported for each method including both mean and standard deviation values over the test dataset (40 slices), and the best mean results are underlined. Yellow arrows indicate a small bone structure, with the counterparts in (f)-(h) emphasized within the black circles.

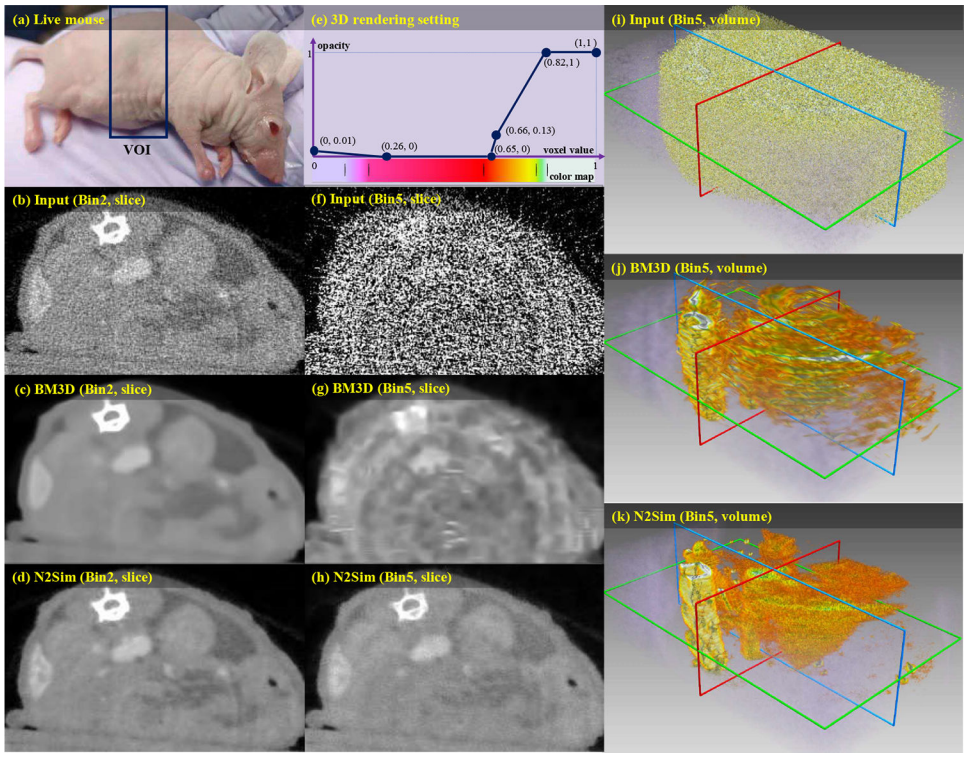

Fig. 7.

Visualization of the denoised PCCT images of a live mouse. (a) shows the live mouse with a boxed area to be scanned. (b), (c), and (d) are respectively the input, BM3D and Noise2Sim results of a slice in the $\mathbf{2}^{\boldsymbol{th}}$ bin. (f), (g), and (h) are the input, BM3D and Noise2Sim results of a slice in the $\mathbf{5}^{\boldsymbol{th}}$ bin. (i), (j), (k) are the 3D rendered images in the $\mathbf{5}^{\boldsymbol{th}}$ bin, where the rendering function of voxel value v.s. opacity and voxel value v.s. color map are given in (e).

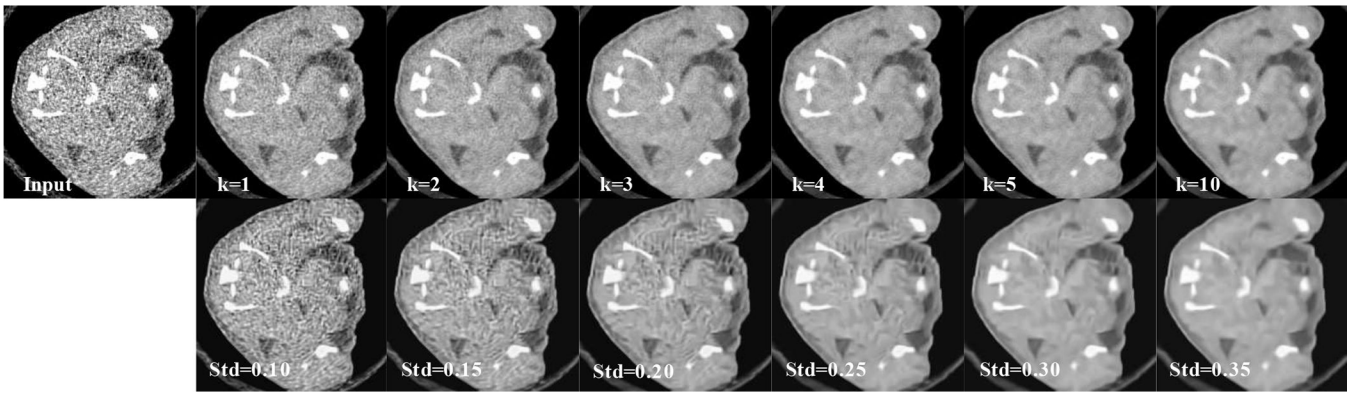

Fig. 8.

Noise2Sim and BM3D results on the died mouse dataset. The first and second rows respectively present Noise2Sim and BM3D results of different denoising levels in the 5^{th} bin.

TABLE I

Quantitative results of different LDCT denoising methods on the Mayo dataset, in terms of mean and standard deviation values over the test dataset (1136 Slices). N2Sim- Means dissimilar pixels are not excluded during training, N2Sim-L1 means using L1 loss, and N2Sim* means directly training the model on the testing dataset. N2Sim is the short form of Noise2Sim.

	Supervised		Weakly-supervised		Unsupervised/Self-supervised					
Methods	**RED-CNN**	**MAP-NN**	**N2Clean**	**CycleGAN**	**N2Void**	**BM3D**	**N2Sim-**	**N2Sim**	**N2sim-L1**	**N2Sim***
PSNR	28.58 ± 1.54	28.28 ± 1.55	**28.78** ± 1.58	23.29 ± 1.11	23.36 ± 1.82	27.28 ± 1.48	27.68 ± 1.31	28.30 ± 1.48	**28.38** ± 1.51	28.33 ± 1.50
SSIM	**90.30** ± 2.92	90.13 ± 2.92	90.21 ± 2.91	87.08 ± 3.20	83.99 ± 4.33	88.30 ± 3.02	89.73 ± 2.73	90.33 ± 2.78	**90.45** ± 2.76	90.37 ± 2.89

TABLE II

The denoising results Using different values of d_{th} in HU. Here all other settings were set to the default values, and L1 loss function was used.

d_{th}	0	10	20	30	40	50	60
PSNR	27.58	28.32	28.38	28.38	28.37	28.37	28.33
SSIM	89.52	90.43	90.48	90.45	90.39	90.32	90.38

TABLE III

The denoising results using different values of $\boldsymbol{S}$ in pixels. Here all other settings were set to the default values, and L1 loss function was used.

d_{th}	3×3	5×5	7×7	9×9	11×11
PSNR	27.10	28.26	28.38	28.34	28.32
SSIM	89.33	90.41	90.45	90.24	90.23

TABLE IV

Generalizability results of different denoising methods on FDA datasets in terms of PSNR and SSIM, with both mean and standard deviation values over the test dataset (408 slices).

Method		Supervised learning			Unsupervised learning		
		RED-CNN	**MAP-NN**	**Noise2Clean**	**Noise2Void**	**BM3D**	**Noise2Sim***
b40f	PSNR	30.37 ± 2.46	30.19 ± 2.30	30.44 ± 2.34	23.43 ± 2.76	27.28 ± 1.28	**30.52** ± 2.12
	SSIM	91.70 ± 3.95	91.28 ± 4.11	91.66 ± 3.88	83.10 ± 5.22	88.31 ± 3.02	**92.03** ± 3.79
b60f	PSNR	19.59 ± 2.43	22.99 ± 0.81	17.71 ± 1.75	17.05 ± 1.02	21.21 ± 2.01	**26.46** ± 0.78
	SSIM	83.36 ± 4.16	81.79 ± 4.79	82.20 ± 0.03	72.42 ± 4.11	80.53 ± 4.95	**91.80** ± 1.28