Computerized Medical Imaging and Graphics 94 (2021) 102010

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Leveraging deep neural networks to improve numerical and perceptual image quality in low-dose preclinical PET imaging

Mahsa Amirrashedi [a,b], Saeed Sarkar [a,b], Hojjat Mamizadeh [a,b], Hossein Ghadiri [a,b], Pardis Ghafarian [c,d], Habib Zaidi [e,f,g,h,*], Mohammad Reza Ay [a,b,**]

[a] Department of Medical Physics and Biomedical Engineering, Tehran University of Medical Sciences, Tehran, Iran
[b] Research Center for Molecular and Cellular Imaging, Tehran University of Medical Sciences, Tehran, Iran
[c] Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran
[d] PET/CT and Cyclotron Center, Masih Daneshvari Hospital, Shahid Beheshti University of Medical, Tehran, Iran
[e] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva CH-1211, Switzerland
[f] Geneva University Neurocenter, Geneva University, Geneva, Switzerland
[g] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[h] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
PET
Small animal imaging
Deep-learning
Low-dose imaging
Denoising

ABSTRACT

The amount of radiotracer injected into laboratory animals is still the most daunting challenge facing translational PET studies. Since low-dose imaging is characterized by a higher level of noise, the quality of the reconstructed images leaves much to be desired. Being the most ubiquitous techniques in denoising applications, edge-aware denoising filters, and reconstruction-based techniques have drawn significant attention in low-count applications. However, for the last few years, much of the credit has gone to deep-learning (DL) methods, which provide more robust solutions to handle various conditions. Albeit being extensively explored in clinical studies, to the best of our knowledge, there is a lack of studies exploring the feasibility of DL-based image denoising in low-count small animal PET imaging. Therefore, herein, we investigated different DL frameworks to map low-dose small animal PET images to their full-dose equivalent with quality and visual similarity on a par with those of standard acquisition. The performance of the DL model was also compared to other well-established filters, including Gaussian smoothing, nonlocal means, and anisotropic diffusion. Visual inspection and quantitative assessment based on quality metrics proved the superior performance of the DL methods in low-count small animal PET studies, paving the way for a more detailed exploration of DL-assisted algorithms in this domain.

1. Introduction

Positron Emission Tomography (PET) has proved its mettle not only in a myriad of clinical applications but also holds great potential in understanding the specific molecular mechanisms that occur in small animal models of human disease. Although the preclinical application of PET imaging is growing by leaps and bounds in the last decades, the amount of injected activity to laboratory animals is the most critical concern hitherto facing translational PET studies, particularly when the repeated administration of radioactive agents to the same animal is required for longitudinal studies. Needless to say, a higher amount of radiotracer guarantees a sufficiently high signal-to-noise ratio (SNR) but may adversely influence the experimental outcomes due to a well-known "mass effect" (Jagoda et al., 2004; Kung and Kung, 2005). This is also extremely important for neuroreceptor studies where the specific activity is problematic and the induced pharmacological effects could violate the tracer principle (Herfert et al., 2020). Indeed, limiting the tracer activity in PET studies not only alleviates the concerns over the toxicity and pharmacological effects of the injected radiotracer but also decreases the amount of radiation to animals and operators in core

* Corresponding author at: Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva CH-1211, Switzerland.
** Corresponding author at: Department of Medical Physics and Biomedical Engineering, Tehran University of Medical Sciences, Tehran, Iran.
E-mail addresses: m-amirrashedi@razi.tums.ac.ir (M. Amirrashedi), sarkar@tums.ac.ir (S. Sarkar), Hmamizadeh@razi.tums.ac.ir (H. Mamizadeh), H-ghadiri@tums.ac.ir (H. Ghadiri), pardis.ghafarian@sbmu.ac.ir (P. Ghafarian), habib.zaidi@hcuge.ch (H. Zaidi), mohammadreza_ay@tums.ac.ir (M.R. Ay).

https://doi.org/10.1016/j.compmedimag.2021.102010
Received 29 June 2021; Received in revised form 25 October 2021; Accepted 26 October 2021
Available online 7 November 2021
0895-6111/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

laboratories, lowers the total imaging cost, improves the scanner throughput, and above all, would set the stage for exploring new generations of radiotracers (Molinos et al., 2019).

It is obvious that low-count PET imaging (small injected activity or fast scan) is characterized by higher levels of noise and thus interpretation inaccuracies. This is of particular concern in animal studies regarding the lower sensitivity of the preclinical scanners and shorter acquisition times due to the compounding effects of anesthesia on animal physiology as well as the rapid metabolism rate in rodents. The situation is further aggregated, for example, in fast dynamic studies, screening short-lived radioisotopes, and multitracer PET investigations, where the number of collected events is inherently limited therein.

Hence, a host of hardware-based and software-based solutions were investigated and deployed to move forward and to resolve the challenges related to count deficiency in low-statistics PET studies (Zaidi and El Naqa, 2021). The former strategies were extensively discussed in (Amirrashedi et al., 2020b, 2020c) and based around developing high-end scanners with improved detection capabilities. While the latter offer more cost-effective solutions and could be segregated into three broad categories: (i) Edge-aware noise attenuating filters (nonlocal means (NLM) (Arabi and Zaidi, 2018, 2020; Buades et al., 2005), bilateral filtering (BF) (Tomasi and Manduchi, 1998), anisotropic diffusion (AD) (Perona and Malik, 1990), and image-guided filters (IGFs) (He et al., 2012), etc.), (ii) iterative reconstruction algorithms (Reader and Zaidi, 2007) (maximum a posteriori (MAP), penalized weighted least-squares (PWLS), total variation based reconstruction, smoothing priors, etc.), and (iii) deep-learning (DL)-inspired methods.

Post-reconstruction filtering and reconstruction-based strategies are yet the most ubiquitous techniques in denoising applications and have a long-lasting history in this domain. However, for achieving state-of-the-art results, the pertinent parameters (e.g., noise level, kernel size, number of iterations, regularization strength, etc.) need all to be adjusted and optimized professionally according to the type of imaging task. Over and above the laborious parameter tuning and frequent disparities resulting from human biases, the pain point is the heavy computational load and time complexity of such algorithms.

The last category belongs to DL-based methods, which are entrusted to learn the ill-posed relation between high-statistics emission images and the low-count counterpart rather than denoising the emission data. From an architectural point of view, Convolution Neural Network (CNN) particularly multi-scale U-Net models along with Generative Adversarial Network (GAN) are already the most preferred types of DL frameworks in the image processing domain. Of late, various custom-built DL paradigms were utilized successfully in a broad range of PET studies, such as noise suppression, sinogram correction, attenuation correction, scattering estimation, low-dose (LD) PET imaging, and direct image reconstruction. One can refer to the most recent reviews on this topic (Arabi et al., 2021; Gong et al., 2020a; Lee, 2021; Reader et al., 2021; Wang et al., 2021; Zaidi and El Naqa, 2021).

In existing works, several authors reported the contribution of modified CNN designs with single or multimodal inputs in the solutions to full-dose (FD) PET estimation. As such, Cui et al. (2019) utilized an unsupervised DL method to recover clean PET images from noisy measurements. To this end, anatomical priors, like CT and MRI images of the patient were considered as the network input, and the noisy images were introduced as the train labels. Xiang et al. (2017) proposed a deep auto-context CNN framework with multimodal input channels to build high-quality standard counting PET images where 2D patches of PET and T1-weighted MRI slices were used for training the network. Similarly, Chen et al. (2019) investigated an adapted U-Net model to gauge high SNR and visually pleasing amyloid PET images by feeding the network with multi contrast MRI scans along with LD PET. da Costa-Luis and Reader (2021) introduced a relatively small but highly efficient CNNs called micro-nets for denoising low-count dataset whereby three realizations of the same PET acquisition provided by different reconstructions and the corresponding T1-weighted MRI image were inputted to the network. This work and similar investigations confirmed that joint multimodal inputs would lead to a significant gain in network performance compared to PET-only models (da Costa-Luis and Reader, 2021). However, besides the inter-modality mismatches and higher imaging costs associated with multiple acquisitions, the application of multimodality-guided networks is limited to hybrid scanners and not feasible with monomodal instruments.

Meanwhile, inspired by the successful implementation of a residual training scheme in sparse-view CT reconstruction and inasmuch as the conventional U-Net structure leads to extremely blurred boundaries, the attention shifted toward the residual U-Net models (Han et al., 2016) and also GAN variants (Gong et al., 2020b; Wang et al., 2018; Zhou et al., 2020). Following this trend, Xu et al. (2017) explored employing a residual U-Net architecture for enabling low-dose brain PET imaging taken at 1/200th of the routine dose. Compared with an auto-context network, NLM, and BM3D filters, residual U-Net presented compelling results in terms of image quality, structural similarity, and execution speed. The authors also proved that supplying the network using 2.5D inputs makes it possible to conserve the contextual information and the tracer uptake pattern more effectively as compared with the single slice training strategy. In a quantitative study conducted on small lung lesions, Lu et al. (2019) demonstrated that a 3D U-Net denoising results in better image quality and lower mean tracer uptake bias compared to Gaussian filtering, CT-guided NLM filter, and MAP reconstruction with quadratic and relative difference priors. Later on, Spuhler and colleagues investigated a 2.5D dilated U-Net with residual and skip connections in LD brain PET restoration. The advantage of a dilated network was shown in predicting sharper and well-restored outputs by removing the down-sampler and up-sampler operators in the conventional U-Net architecture (Serrano-Sosa et al., 2020; Spuhler et al., 2019). Another study on LD to FD mapping was performed by Sanaat et al. in which the author proved the qualitative and quantitative superiority of sinogram-domain training compared to image-domain FD recovery, though at the cost of 6-fold larger computation time (Sanaat et al., 2020). This work was extended for whole-body (WB) PET imaging using modified cycle-consistent generative adversarial network (CycleGAN) and residual neural network (ResNet) models (Sanaat et al., 2021). Lei et al. (2019) unveiled the superior performance of GAN variants in WB PET denoising. In a recent study by Gong et al. (2020b) it was demonstrated that a Wasserstein GAN (WGAN) framework trained with mixed adversarial loss function has higher denoising performance in contrast to networks trained with pure adversarial or conventional objective functions.

Albeit being extensively examined for clinical purposes, so far as we know, no study has yet to explore the feasibility of DL-based image enhancement in the low-count preclinical scenarios. As the case in humans, the administered dose and acquisition time are standing as the most challenging hurdles in murine studies. Motivated by the diverse applications of DL techniques in clinical trials and owing to its simple implementation and phenomenal success, herein, we examined different DL-based frameworks to generate FD small animal PET images from undersampled scans mimicking actual LD scanning conditions. We also compared the results of DL-based FD reconstruction with other workhorse filters established for denoising applications; including Gaussian filter (GF), block-wise median-guided NLM (BMNLM), and anisotropic diffusion filter (ADF). To the best of our knowledge, this is the first study utilizing DL techniques to synthesis high-quality small animal PET images using 20% of the recorded counts. Our main motivation is decreasing the administered activity as well as imaging time in the dedicated Xtrim-PET scanner (Amirrashedi et al., 2019) in which WB mice imaging necessitates acquiring two bed positions, doubles the scan duration, and adversely disturbs the uptake pattern to that of WB scanners.

2. Materials and methods

2.1. Image acquisition and dataset preparation

Prior to initiating the study, ethical clearance was obtained from the Animal Care and Ethics committee of the Tehran University of Medical Sciences (Approval number: IR.TUMS.MEDICINE.REC.1397.004). All procedures were conducted in a small animal core facility. Throughout the experiments, animals had enough access to water/food and were continuously monitored by professionally trained personnel. The small animal PET images for this study were collected through a high-resolution Xtrim-PET camera dedicated to rodent imaging (Amirrashedi et al., 2019). Our dataset consists of fourteen mice (8 females/6 males) with an average weight of 32 ± 10 gr. Animals have received a dose of 11 ± 2 MBq ^{18}F-FDG and were kept on a heating pad during the uptake period. Five minutes prior to scan initiation, anesthesia was performed by intraperitoneal administration of a ketamine/xylazine mixture (100 mg/kg ketamine + 10 mg/kg xylazine). About 45–60 min after tracer injection, each subject was scanned under the scanner default protocol settings (energy window = 350–650 keV, timing window = 5 ns, FOV = 80 mm, 2 bed positions, 10 min per bed).

For each acquisition, 1/5th of the counts were selected randomly to emulate the low-count acquisition whereas the normal-count data were generated by considering all coincidences detected during the scanning period (Schaefferkoetter et al., 2019). We have examined the reliability of the method through several phantom experiments, some of which were described in the next section. To increase the number of samples and to compare different methods, we synthesized five realizations for each subject. To this end, we randomly downsized the FD counts into five independent LD list mode files, each containing 20% of the events. All images were reconstructed through an ordered subsets expectation maximization (OSEM) algorithm (5 iterations, 8 subsets by considering decay, normalization (Amirrashedi et al., 2020a)), attenuation (two-level segmentation method), and scatter corrections with tail fitting. There were 121 slices per scan and reconstructed images had 130 × 130 pixels with voxel sizes of 0.615 mm × 0.615 mm × 1.05 mm. All slices containing the thorax and abdomen regions were selected to train the network (from the neck to the lower abdomen).

2.2. Synthetic LD generation

In this study, we approximated the LD conditions (synthetic LD) by randomly down-sampling the events collected during the standard-dose acquisitions. From a theoretical point of view, the overall image quality should be identical for the emission images reconstructed from the same number of true events. Therefore, it is acceptable to use down-sampled LD images as a surrogate for the actual LD conditions performed in the same count levels (Schaefferkoetter et al., 2019). However, one should take into account the rate of true and random coincidences, which strongly depends upon the activity concentration and the object size within the FOV. In this section, we describe two phantom experiments to showcase the validity of the hypothesis.

In the first experiment, we filled the NEMA image quality (IQ) phantom with an activity concentration of 7.5 μci/gr and scanned it for 10 min. In the second study, a uniform cylinder with a 1.5 cm diameter and 6.5 μci/gr was placed inside the FOV and imaging was performed for 5 min. The cylinder was located inside a water phantom with a 3 cm diameter. For both studies, the acquisition was repeated until the activity level in the phantom reached 0.2 of its initial value (actual LD scans). To generate synthetic LD data from the standard-dose scans, we randomly down-sampled the 20% of the events stored in the list mode files during FD acquisition. All data, including FD, synthetic LD, and actual LD images were corrected (for attenuation, normalization, decay, scattering, and random) and reconstructed using OSEM algorithm (5 iterations, 8 subsets).

2.3. The network architecture

2.3.1. U-Net models

We trained four networks based on a multi-scale U-shaped architecture tweaked for our purpose (Ronneberger et al., 2015). The base model is illustrated in Fig. 1 and implemented in Keras and Tensorflow libraries (Abadi et al., 2016; Chollet, 2018). Our modified U-Net model is composed of an analysis module/encoder to extract the high-level features by compressing the input image, and a synthesis module/decoder to construct the labels by mapping the extracted features into the full dataset representation. Both encoding and decoding subnetworks include three steps with subsequent implementation of two convolutional blocks (shown in blue). Each block is formed by a convolutional layer with a kernel width of 3, followed by a batch normalization (BN) and an Exponential linear unit (ELU). Compared to the rectified linear unit (ReLU) and its variants, ELU is less sensitive to noise. Therefore, we chose ELU as the activation function for training procedures (Clevert et al., 2015). Downsampling of the feature map is accomplished using an strided convolution in the encoder side whilst to preclude the checkboard artifacts (which was severe in our case), a bicubic interpolation algorithm is performed to increase the dimensionality in the decoding step. To avoid resolution loss and to prevent gradient diffusion, concatenating connections are also integrated to pass the same-scaled features from the encoder to the corresponding stage in the decoder. Finally, a convolution layer with a kernel width of 1 is utilized to map the feature vector to the network output.

2.3.2. GAN framework

The proposed GAN framework has two main components: 1) a U-Net-like generator G which takes low-dose PET and predicts its full-dose counterpart 2) a discriminator D which tries to distinguish between predicted images by the generator (PD) and the true full-dose images (FD). Instead of using a simple CNN model, the modified U-Net model (Fig. 1a) was used as the generator for the GAN model used in this study.

The second part of the network is the discriminator (Fig. 1b) which includes four strided convolution layers with 3 × 3 kernels and two fully connected layers. We used BN and LeakyReLU after all convolutions. The number of kernels is 32/64/128/256/128/64/32/1 for the generator and 64/128/256/512/1024/1 for the discriminator.

2.4. Training strategies and hyperparameters

2.4.1. U-Net training

Given the challenges imposed by volumetric training models (e.g., large memory footprints, time, and computation constraints), we opt to train the networks on 2D and 2.5D (called synthetic 3D) schemes with the same encoder-decoder architecture therein before discussed. The most notable difference between 2D and 2.5D concepts is the input layer which could be defined as a vector with a size of (C × S × W × H), where C is the number of channels, S indicates the number of slices along the z-direction, and (W × H) denotes the in-plane dimensions for each transaxial slice. According to the above definitions, (1 × 1 × 128 × 128) and (3 × 3 × 128 × 128) vectors were inputted for training 2D and 2.5D models, respectively.

Our 2.5D set-up takes the 3 juxtaposed slices (the source slice coupled to the superior/inferior slices) as different input channels to predict the output for the middle slice. Doing so provides more context information and spatial cues comparing to its single slice 2D alternative, empowers the U-Net to differentiate between radiotracer uptake patterns and noise texture, yielding better results.

For both 2D and 2.5D models, we compared the image-to-image transformation to that of residual mapping. In residual networks, mapping relies on the difference between full and low-count images in lieu of directly estimating the original FD images. Since residual learning is easier than direct mapping, convergence is faster.

As summarized in Fig. 2(a)–(d), we trained four separate U-shaped

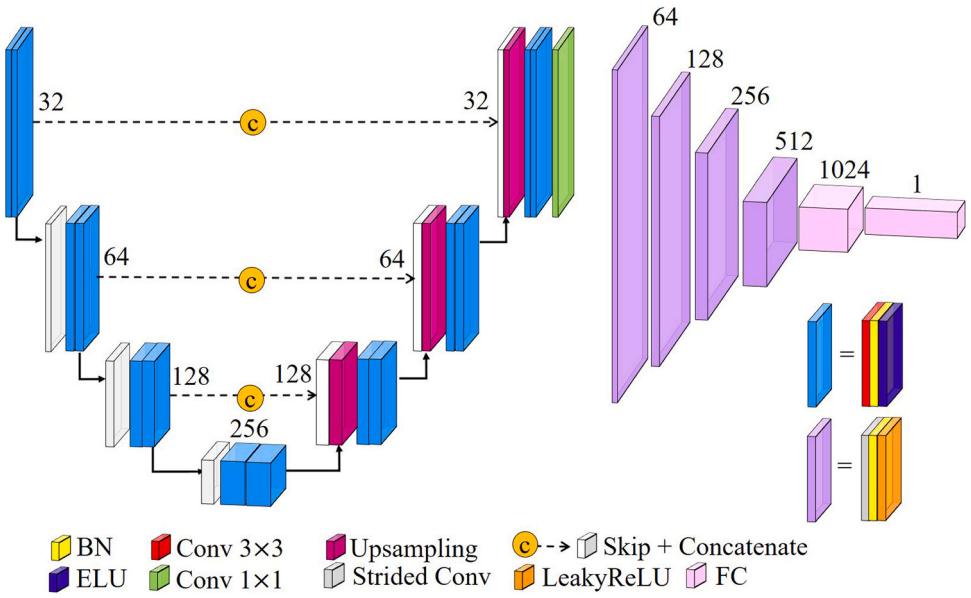

Fig. 1. Schematic illustration of (a) proposed U-shaped model and (b) discriminator. The U-shaped model was used as the generator in the PTWGAN framework.

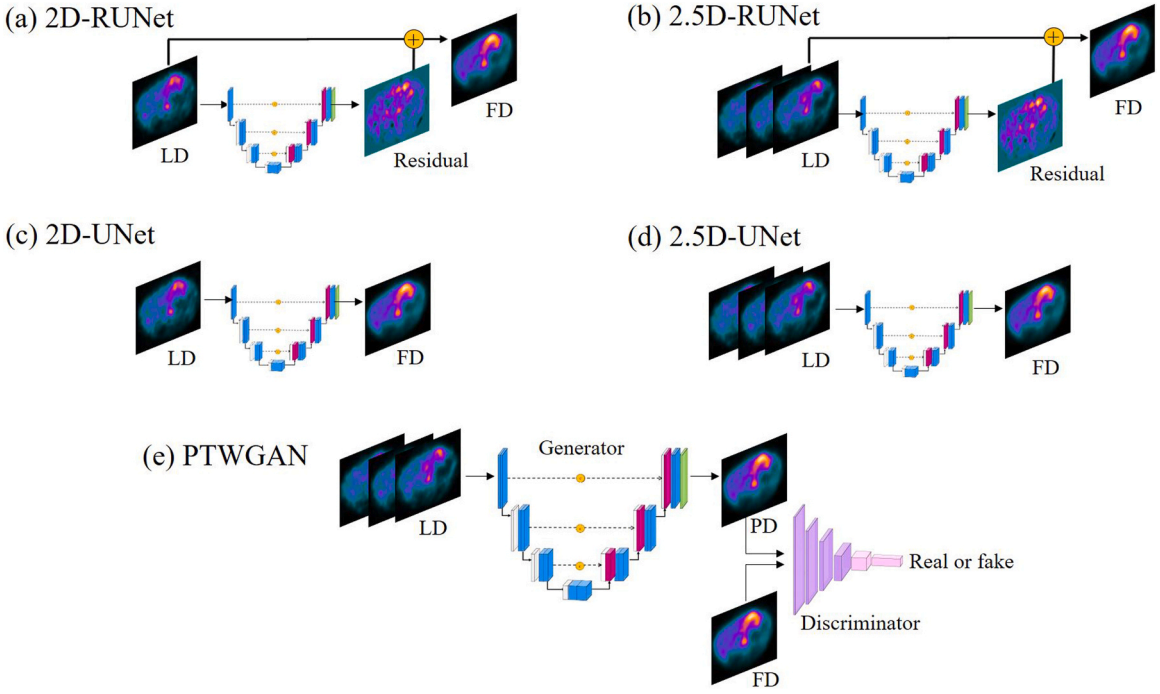

Fig. 2. Training strategies in this study. (a) 2D-RUNet: the model takes the target slice as input and predicts the corresponding residual map. (b) 2.5D-RUNet: The model takes three adjacent slices as input and predicts the residual map for the middle slice. (c) 2D-UNet: the model takes the target slice as input and directly predicts the FD image. (d) 2.5D-UNet: the model takes three adjacent slices as input and predicts the FD image for the middle slice. (e) PTWGAN: the generator takes three slices and predicts the middle one which then goes through the discriminator.

models: 2D-UNet (single slice LD/FD mapping), 2D-RUNet (single slice LD/residual mapping), 2.5D-UNet (multi-slice LD/FD mapping), and 2.5D-RUNet (multi-slice LD/residual mapping).

The models have trained over 200 epochs (there were no further improvements noticed in the model) with a mini-batch size of 10 and using the Adam optimization approach with a learning rate (LR) of 1e − 4. The mean absolute error (MAE) was used to compute the loss between the network prediction and target-truth label since it is more robust to noise and artifacts in comparison to mean squared error (MSE) (da Costa-Luis and Reader, 2021; Wang et al., 2018; Xu et al., 2017). To improve the network generalizability, 4-fold augmentation was performed and the number of samples was enhanced by randomly rotating, translating, and flipping (horizontally and vertically) the data during the training phase.

2.4.2. *Training WGAN based on transfer-learning*

Following the latest studies (Gong et al., 2020b; Shan et al., 2018), we trained the proposed GAN model using Wasserstein distance and improve it by adding a gradient penalty (Eq. (1)). Gong et al. (2020b) introduced a hybrid 2D and 3D as the denoising model while Shan et al.

(2018) transferred the parameters of a pre-trained 2D Conveying Path-based Convolutional encoder-decoder to initiate the generator with 3D kernels in a PTWGAN framework. Herein, we preferred to use the parameters of the previously trained 2.5D design to train our PTWGAN model. Furthermore, due to the noisy nature of the LD images, we added the MAE term (instead of MSE term in (Gong et al., 2020b)) to the training loss of the generator to decrease the amount of noise while enhancing the pixel-wise similarity between PD and FD images (Gong et al., 2020b; Shan et al., 2018). The adversarial loss functions used for training the proposed PTWGAN is summarized in (1). The generator and discriminator of the proposed network were optimized following (2) and (3), respectively. W_{gp} and W_{MAE} are the weighting parameters for the gradient penalty and MAE terms in the mixed loss function, respectively.

$$\begin{aligned}\mathscr{L}_{adv} &= \underbrace{\mathbb{E}_{FD}[D(FD)] - \mathbb{E}_{LD}[D(PD)]}_{wasserstein\ distance} \\ &\quad + W_{gp} \times \underbrace{\mathbb{E}_{I}[\|\nabla D(I) - 1\|^2]}_{gradiant\ penalty}\end{aligned} \tag{1}$$

$$\mathscr{L}_G = -\mathbf{E}_{LD}[D(PD)] + W_{MAE} \times \mathscr{L}_{MAE} \tag{2}$$

$$\begin{aligned}\mathscr{L}_D &= \mathbb{E}_{LD}[D(PD)] - \mathbb{E}_{FD}[D(FD)] \\ &\quad + W_{gp} \times \mathbb{E}_{I}[\|\nabla D(I) - 1\|^2]\end{aligned} \tag{3}$$

$$I = \beta \times PD + (1 - \beta) \times FD \tag{4}$$

$\mathbb{E}_b[a]$ is the expectation of a as a function of b,∇ is the gradient operator, in (4) I is the sample synthesized between real and fake images in which β has a uniform distribution in the interval of [0,1]. As suggested in Gong et al. (2020b), task-specific initialization for the generator decreases the difficulty of the training procedure and enhances the performance of the network in low-dose imaging. To this end, first we trained a 2.5D U-Net model directly from the scratch with MAE loss, Adam optimizer (LR = 1e – 4), and epochs = 40. In the second step we transferred the parameters of the trained model to initialize the generator in the WGAN framework. Similar to previous studies (Gong et al., 2020b; Shan et al., 2018), we trained the generator and discriminator separately where the number of extra training steps for the discriminator was 4. For training based on transfer learning, we used the following settings for hyperparameters: Adam optimizer (LR = 5e−5), batch-size = 40, epochs = 10.

As has been suggested (Shan et al., 2018), we set the weighting parameter for gradient penalty to 10, but for finding the optimum MAE weighting parameter we examined different values for W_{MAE}. Finally, we achieved the lowest NRMSE on test data when the weighting factor for the MAE term was 10^5 (see Fig. 3).

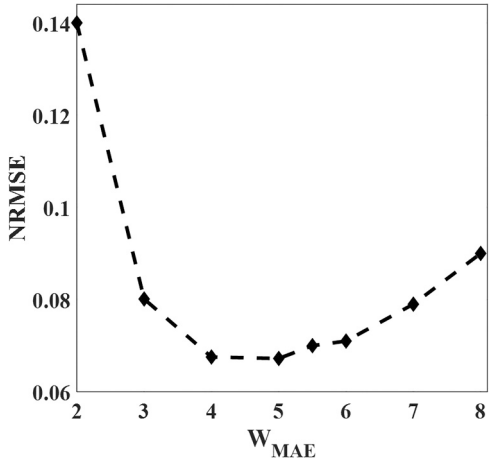

Fig. 3. The effect of the MAE term in PTWGAN on NRMSE. Values were calculated for the external test dataset.

2.5. Cross-validation

In this work, we considered two different scenarios. In the former, we investigated different DL models for our particular case. In this regard, we selected three WB mice scans for testing the final model (not included in the training and validation phase) and used the rest for training and hyperparameter tuning.

In the second scenario, we conducted leave-one-group-out cross-validation to further appraise the network performance and to compare different post-reconstruction denoising methods as suggested in [23,27]. For 14 mice, we trained the network with 6 groups of mice (12 mice) and tested on the reminder group (2 mice). We repeated this process 7 times to retrieve the FD images for each mouse. After training the network, low-count test images (considering 5 realizations for each mouse) were inputted into the model to synthesis the corresponding labels.

It should be noted that 16,800 axial slices were used for training of the proposed model, including 12 (number of mice) × 70 (number of axial slices for each mice) × 5 (number of realizations generated for each sample) × 4 (number of augmentations over a single image). The test data set consisted of 700 axial slices, including 2 (number of mice) × 70 (number of axial slices for each mice) × 5 (number of realizations generated for each sample).

2.6. Comparison to existing filtering approaches

As per discussed, we investigated the competitive performance of the DL method against the existing filtering techniques including Gaussian post-reconstruction filtering, block-wise median-guided NLM [37], and anisotropic diffusion as well [10].

2.6.1. Gaussian filter

The smoothing strength of the GF depends solely upon the standard deviation of the kernel. To preclude the over smoothing effect and subsequent resolution loss, the standard deviation was set to 0.6.

2.6.2. Block-wise median-guided NLM filter

Being one of the most powerful noise cancellation concepts, the classical NLM formula replaces each voxel in the noisy image with a weighted average of all voxels in a predefined search window centered at the target voxel v [6]. The block-wise implementation allows computing the NLM solution in a 3D manner rather than 2D patches. The weight between two voxels v and v' is then calculated by taking to account the similarity of their neighborhood configuration. The filtering steps could be expressed as:

$$I_{BMNLM}(v) = \frac{1}{NF(v)} \sum_{v' \in SW} w(v, v')\, LD(v') \tag{5}$$

$$M(v) = Median(v, K) \tag{6}$$

$$W\left(v, v'\right) = \exp\left(-\frac{\|M(v,s) - M(v',s)\|}{h^2}\right) \tag{7}$$

With more detail, I_{BMNlM} is the smoothed image evolved by BMNLM filter, NF is a normalization constant to guarantee that the summation of the weights for the target voxel equals 1, M is the filtered LD image using a median filter with a kernel size of K, and $w\ (v, v')$ stands for BMNLM weight between two voxels of v and v', $\|.\|$ is the L2-norm between the similarity patches/blocks of v and v', s is the box-shaped similarity window, h specifies the decay of the exponential function and thus controls the smoothing performance of the filter. It is determined based on the standard deviation in a homogeneous background region. To avoid over-weighting of the target voxel and to take care of information-bearing features, we implemented an extended version of the BMNLM as suggested by Chan et al. [37]. To this end, we first applied a 3D median kernel with a small size of 3 on the noisy LD images to remove the

transient speckle noise and then extracted the BMNLM weights from the median filtered images. The estimated weights were then implemented on the noisy LD image to achieve the final results. Note that median filtering is only used to extract the filter weights. As a tradeoff between computational cost and the filtering performance, a search window of $5 \times 5 \times 5$ and a similarity window of $3 \times 3 \times 3$ were chosen empirically for all the experiments.

2.6.3. Anisotropic diffusion filter

ADF is another remedy toward edge-preserving noise cancelation, wherein denoising operation is being modeled as an iterated diffusive process, encourages the diffusion in regions with small inter-pixel fluctuations while avoiding in areas with high variations (e.g., boundaries, edge, prominent features). Mathematically, the AD equation is formulated as:

$$\begin{cases} \dfrac{\partial I_{ADF}(v,t)}{\partial t} = div[g(\|\nabla LD\|)\nabla LD] \\ I_{ADF}(v,0) = I_{0ADF}(v) \end{cases} \tag{8}$$

$$g = -\exp\left(\frac{|\nabla LD|}{\kappa}\right)^2 \tag{9}$$

I_{ADF} is the image restored using the ADF equation, g is the conduction coefficient, which depends on the image gradient in different directions, div is the divergent operator, and t is used to enumerate the iterations. We used the quadratic model in (9) to compute the conduction coefficients since it gives better results in terms of contrast. Using the 3D version of the ADF with 26 neighbors, we set the value of diffusion rate to $< 3/44$ to keep the stability of the diffusion process. Following [38], we selected the edge-conservation term κ empirically according to the standard deviation inside a homogeneous ROI.

2.7. Quantitative analysis

We investigated the restoration performance of the proposed networks along with other denoising strategies, in terms of normalized root mean square error (NRMSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM), and normalized cross-correlation (NCC) defined as follow:

$$SSIM(x,y) = \frac{(2\sigma_{xy} + C_2)}{\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \times \frac{(2\mu_x\mu_y + C_1)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)} \tag{10}$$

$$NRMSE\left(I_E, I_G\right) = \sqrt{\frac{\sum_{v=1}^{V}(I_E(v) - I_G(v))}{\sum_{v=1}^{V}\left(I_G^2(v)\right)}} \tag{11}$$

$$PSNR\left(I_E, I_G\right) = 20 \times \log_{10}\left(\frac{MAX_G}{\sqrt{MSE(I_E, I_G)}}\right) \tag{12}$$

$$NCC(I_E, I_G) = \frac{\sum_{v=1}^{V}\left(\left(I_G(v) - \bar{I}_G\right)\left(I_E(v) - \bar{I}_E\right)\right)}{std(I_G) \times std(I_E)} \tag{13}$$

I_G refers to the ground-truth corresponding to the FD image in our case. The image under evaluation is indicated by I_E which is either the LD image or the restored image following one of the methods described in previous sections. In (10), μ_x and μ_y show the mean intensity value for I_E and I_G, respectively. C_1 and C_2 are the constants to avoid a null denominator. σ_{xy} is the covariance of $I_E$and I_G,whereas σ_x and σ_y indicate the variances. V indicates the number of voxels inside the body, MAX_G is the maximum value in I_G and MSE presents the mean squared error between I_E and I_G.

In (13), $\bar{I}$and $std(.)$ compute the mean and standard deviation inside the masked volume, respectively. To get more realistic results and to exclude the background voxels, a binary mask was applied to the WB PET images. Image quality metrics were then computed over all the voxels that lie inside the body volume. We also performed a joint histogram analysis to illustrate the voxel-wise correlation of the standardized uptake values (SUV) between denoised and reference FD PET images.

Additionally, we computed the coefficient of variation (CV) and recovery coefficient (RC) to compare the synthetic LD and actual LD phantom scans. For the uniform region, the mean activity concentration and standard deviation were calculated inside a 3D volume of interest (VOI) encompassing 70% of the phantom size. For hot spheres inside the IQ phantom, we calculated the average of the voxel intensities $\geq$ 80% of the maximum voxel. CV was defined as the ratio of standard deviation to the mean value inside the predefined VOI, and the RC was the ratio of mean activity in the VOI to the true activity inside the phantom.

2.8. Statistical analysis

For statistical analysis, a paired-sample t-test was carried out in Prism 8 (Graph Pad Software Inc., San Diego, CA, USA) to compare the image quality metrics (PSNR, NRMSE, NCC, and SSIM) among different methods. Bonferroni adjusted P-values less than 0.05 were deemed to reflect statistical significance. All the metrics were computed relative to the target-truth image. Moreover, when comparing different denoising methods, evaluation metrics were calculated across five realizations for each subject and reported as mean $\pm$ standard deviation.

3. Results

3.1. Comparison between synthetic and actual LD scans

Image quality metrics for FD, synthetic LD, and actual LD scans for IQ and uniform phantoms were presented in Table 1. Transverse slices from IQ phantom experiment were also illustrated in Fig. 4. The image characteristics are almost identical between down-sampled and actual LD conditions (P-value > 0.05) which shows that the synthetic LD images used in this work are appropriate surrogates for actual LD scans (this holds for the activity concentrations and object size used in this work).

Table 1
Image quality metrics for FD, synthetic LD, and actual LD scans.

NEMA IQ phantom		**FD**	**Synthetic LD**	**Actual LD**
NRMSE		–	0.113	0.111
PSNR		–	33.71	33.76
NCC		–	0.981	0.986
SSIM		–	0.879	0.881
Uniform region	RC	1.051 ± 0.06	1.048 ± 0.09	1.01 ± 0.08
	CV	5.7%	8.5%	7.9%
Hot rods (diameter)	RC (5 mm)	0.742 ± 0.1	0.72 ± 0.17	0.733 ± 0.16
	RC (4 mm)	0.654 ± 0.1	0.656 ± 0.16	0.667 ± 0.16
	RC (3 mm)	0.449 ± 0.08	0.447 ± 0.12	0.448 ± 0.11
	RC (2 mm)	0.327 ± 0.09	0.315 ± 0.1	0.313 ± 0.09
Uniform phantom		**FD**	**Synthetic LD**	**Actual LD**
NRMSE		–	0.108	0.096
PSNR		–	36.9	37.1
NCC		–	0.985	0.987
SSIM		–	0.973	0.975
Uniform region	RC	0.95 ± 0.072	0.93 ± 0.08	0.92 ± 0.075
	CV	7.8%	8.6%	8.1%

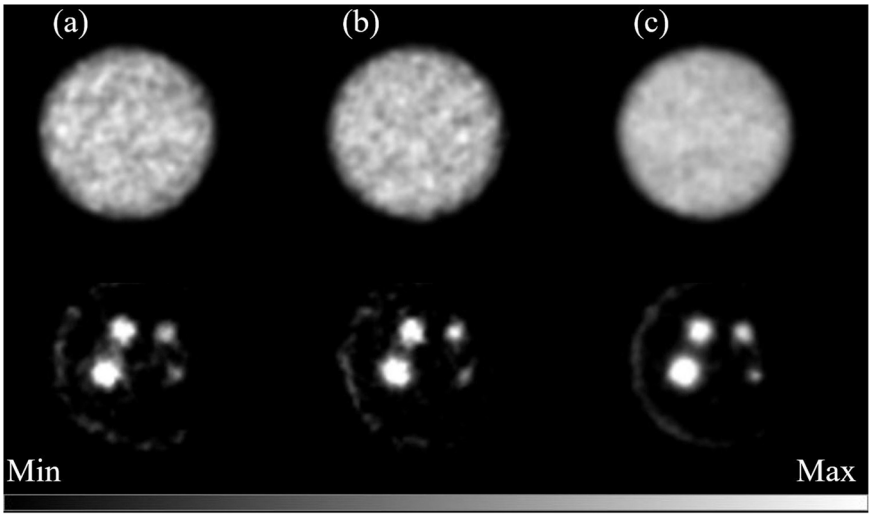

Fig. 4. Representative PET images of the NEMA IQ phantom from uniform section (top) and hot rods (bottom). From left to right: a) actual LD, b) synthetic LD, c) FD images.

3.2. Comparison among DL models

3.2.1. U-Net models

Fig. 5 illustrates several examples for a 38 gr mouse PET scan from different (coronal, sagittal, and transverse) views. From left to right, images represent (a) noisy LD images, synthesized FD images with (b) 2D-UNet, (c) 2D-RUNet, (d) 2.5D-UNet, (e) 2.5D-RUNet, along with the corresponding (f) ground-truth or FD image. From the data in Fig. 5, it is apparent that irrespective of the model, DL can significantly reduce the amount of noise in the inputs images without sacrificing the key information and organ boundaries. Moreover, shape distortions in the myocardium and the fuzzier appearance of the intestinal tract are markedly recovered through the DL pipelines. However, the images treated with the 2D configuration exhibit pixelated appearance and stair-step artifacts (pointed by red arrows), particularly in the coronal and sagittal views, indicating the paucity of information along the axial direction. The same results are also reported in clinical applications of DL-based denoising (Lu et al., 2019). At the other extreme, the images synthesized through the 2.5D concept are more uniform, better decode the uptake pattern using the volumetric information coming from adjacent slabs, and thus yielding comparable qualities to the original FD images. The differences between 2D and 2.5D models are therefore more discernable for the organs with contiguous uptake pattern that encompasses multiple axial planes including paraspinal regions (marked in red) or gastrointestinal (GI) tract (annotated by green arrows).

In comparison to 2.5D-UNet, 2.5D-RUNet shows slightly better performance at distinguishing the noise content from the sharp boundaries. It could also be inferred that 2.5D-RUNet ensures numerically better predictions when comparing the myocardium intensities (blue arrows) and the intestine tract (green arrows) between 2.5D models. The same results could be found when looking closer at the images subjected to 2D networks. Indeed, 2.5D-RUNet suppresses the noise using the

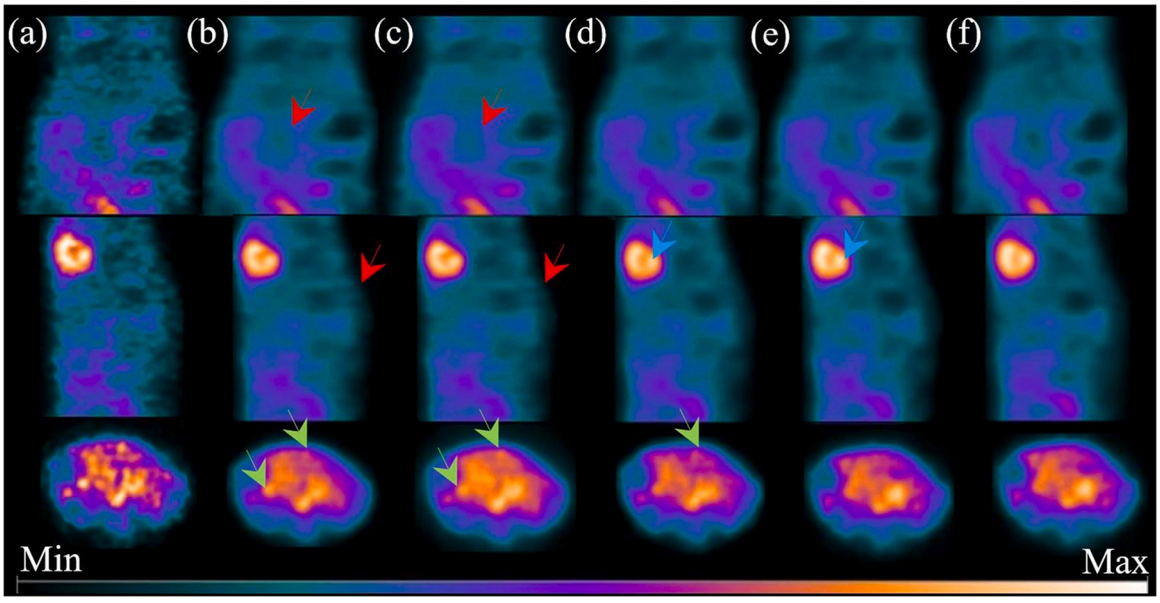

Fig. 5. Comparison among U-Net models. From top to bottom: representative PET images of a mouse showing coronal, sagittal, and transverse views from abdominal section. From left to right: (a) noisy LD images, (b-e) predicted images with 2D-UNet, 2D-RUNet, 2.5D-UNet, 2.5D-RUNet respectively, (f) original FD image (ground-truth). Green arrows: overestimations in gastrointestinal tract, Blue arrows: myocardium, Red arrows: stair-steps artifacts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

information provided by the multiple input channels and grasps the high-frequency contents conveyed by the strong edges through a direct connection with the input image.

Quantitative comparisons among the models are also tabulated in Table 2. Significant differences across all quality metrics were found between the LD images and those obtained with DL models (P-value < 0.01). Supporting the results obtained from the visual interpretations, 2.5D networks beat their 2D counterparts in terms of NRMSE, PSNR, and structural similarity, quantifying the fact that the 2.5D setting results in high-quality predictions while preserving the numerical consistency of data. However, a barely meaningful difference was found between 2.5D models.

3.2.2. U-Net vs. PTWGAN

When comparing to pure U-shaped models, mixed adversarial loss functions results in denoised images with better textural information (Gong et al., 2020b; Shan et al., 2018). Fig. 6 demonstrates the image predicted by 2.5D-UNet with pure MAE loss and the output of the generator trained through mixed adversarial objective functions.

Images treated with PTWGAN better resemble the true FD images in appearance and texture (see the green arrows indicating duodenum and jejunum in the GI tract). Supporting the results reported in (Gong et al., 2020b; Lu et al., 2019), although training with pure MAE loss function leads to slightly smaller NRMSE and higher PSNR values (Table 2) outputs from PTWGAN have a closer texture pattern to the original FDs. Moreover, PTWGAN preserves more detailed information particularly in organs with inhomogeneous tracer uptake like liver and pancreas (yellow arrows), intestine (marked in green), and kidneys (white arrows).

It is worth noting that, we also examined pure adversarial loss (without additional MAE term) and the results were unsatisfactory (not shown here) on account of severe edge artifacts which were also reported in clinical applications of PTWGAN (Gong et al., 2020b). Considering its superior performance (comparing to other U-Net models) and easy training scheme (comparing to PTWGAN), the 2.5D-RUNet was chosen as the baseline for the remaining of our study.

3.3. Comparison with other denoising methods

In this section, we compared the outcomes achieved with 2.5D-RUNet against other post-reconstruction filtering methods with 7-fold cross-validations. The visual comparison among different restoration methods is provided in Fig. 7 & 8. Upon visual inspection, all of the denoising methods lead to noticeable improvements in image quality and allow a significant reduction of the noise level in comparison to the LD image, however, the improvement rate is different among methods. Expectedly, smoothed images passed through a GF kernel are characterized by blurred boundaries around the liver and uterus horn (yellow arrows) whilst with other denoising methods, better performance was obtained in terms of edge-preservation, indicating optimal parameters setting in this study. Even though a preprocessing step is executed before computing ADF and BMNLM weights, images subjected to the ADF algorithm are prone to stair-case artifacts when dealing with high-frequency speckles or sudden alterations in intensity values (e.g., white arrows around uterus horn and red arrows pointing paraspinal region in Fig. 7). These artifacts are generated when the amplitude of the noise is close to the signal intensity. The BMNLM filter presents a more uniform denoising performance against ADF and GF, but the smoothed images slightly have a patchy-like appearance.

Table 2
Quantitative comparison among Different DL networks. Quality metrics were calculated for test dataset and reported as Mean ± standard deviation.

	NRMSE	PSNR	NCC	SSIM
LD	0.106 ± 0.001	35.2 ± 0.6	0.96 ± 0.0025	0.958 ± 0.006
2.5D-RUNet	0.069 ± 0.002	37.13 ± 0.67	0.987 ± 0.0019	0.984 ± 0.003
2.5D-UNet	0.07 ± 0.0006	36.97 ± 0.58	0.986 ± 0.0005	0.982 ± 0.002
2D-RUNet	0.077 ± 0.001	36.76 ± 0.53	0.983 ± 0.0007	0.976 ± 0.004
2D-UNet	0.085 ± 0.003	36.05 ± 0.61	0.981 ± 0.002	0.973 ± 0.007
PTWGAN	0.071 ± 0.0009	36.89 ± 0.6	0.987 ± 0.0007	0.982 ± 0.003
P-value				
(2.5D vs. 2D)-RUNet	**0.002**	**0.007**	**0.0013**	**0.0089**
(2.5D vs. 2D)-UNet	0.093	0.069	**0.0064**	**0.0048**

When compared to traditional filtering, images generated by DL are more homogenous in all views. It could be seen that DL attenuates the unwanted noise more than high contrasted subtle details with the advantage of producing artifact-free volumes that more closely resemble the original ones. The difference between the DL denoising and alternative approaches is more pronounced for the organs housed within the thoracic cage which could be attributed to a higher amount of noise in the mediastinal section. Furthermore, better shape recovery, for example, in the heart (blue arrows in Fig. 8), uterus horn (white arrows in Fig. 7), and the vertebral column was assessed through the DL whilst these regions were ignored by the rival methods. Similarly, the same results could be deduced for the lower body section where the activity in the small intestine and colon (green arrows in Fig. 7 & 8) was preserved more efficiently with DL, reflecting the inherent capability of neural networks to learn the relation between LD and FD pairs rather than purely denoising the input images. Quantitative data averaged on fourteen mice studies are summarized in Table 3. Moreover, for each subject, NRMSE, PSNR, SSIM, and NCC plots are illustrated in Fig. 9, whereby mean and standard deviations are calculated across five realizations for each subject. The normalized root mean squared error averaged on WB PET scans of fourteen mice was 0.114 ± 0.016, which decreased to 0.073 ± 0.009 with 2.5D-RUNet, compared to 0.085 ± 0.011, 0.079 ± 0.008, and 0.081 ± 0.007 after post-processing with GF, ADF and BMNLM, respectively. Similarly, PSNR value improved from 34.76 ± 2.1 dB on LD images to 36.71 ± 1.94 dB, 36.1 ± 1.97 dB, 36.41 ± 1.98 dB, and 36.28 ± 1.8 dB with 2.5D-RUNet, GF, ADF, and BMNLM, respectively. Structural similarity between synthesized images and corresponding original PET scans was also notable for all the methods. Average SSIM and NCC were 0.962 ± 0.008 and 0.965 ± 0.008 for LD images, respectively. Both metrics increased up to 0.986 ± 0.003 for the images enhanced with the DL model. Respective values were 0.98 ± 0.004 and 0.978 ± 0.004 with GF, 0.982 ± 0.003 and 0.981 ± 0.004 with ADF, and 0.981 ± 0.004 with BMNLM methods.

From the joint histogram analysis and linear regression plots (Fig. 10), a higher correlation and voxel-wise correspondence can be observed between DL predicted and FD pairs ($R^2 = 0.986$) compared to other denoising methods ($R^2 = 0.966$ for GF, $R^2 = 0.98$ for ADF, and $R^2 = 0.977$ for BMNLM). In addition, a relatively lower correlation ($R^2 = 0.944$) wasachieved between LD and FD images, indicating poor image quality and large local bias in SUV values.

4. Discussion

The main thrust of this study is to generate preclinical PET images of standard quality from only 1/5th dose PET scans using a fully automated pipeline. Our findings confirm the prior works in the context of DL-based full-dose recovery. However, we pioneered to explore the application of DL-enabled image synthesis in the realm of preclinical practices where the importance of dose minimization is more so than the clinic. This is because the maximum injectable volumes in rodents are considerably smaller compared to that of humans. Limiting the amount of tracer dose in preclinical imaging is manifold where reducing the radiation-induced hazards is only one benefit. Small injected activities circumvent the

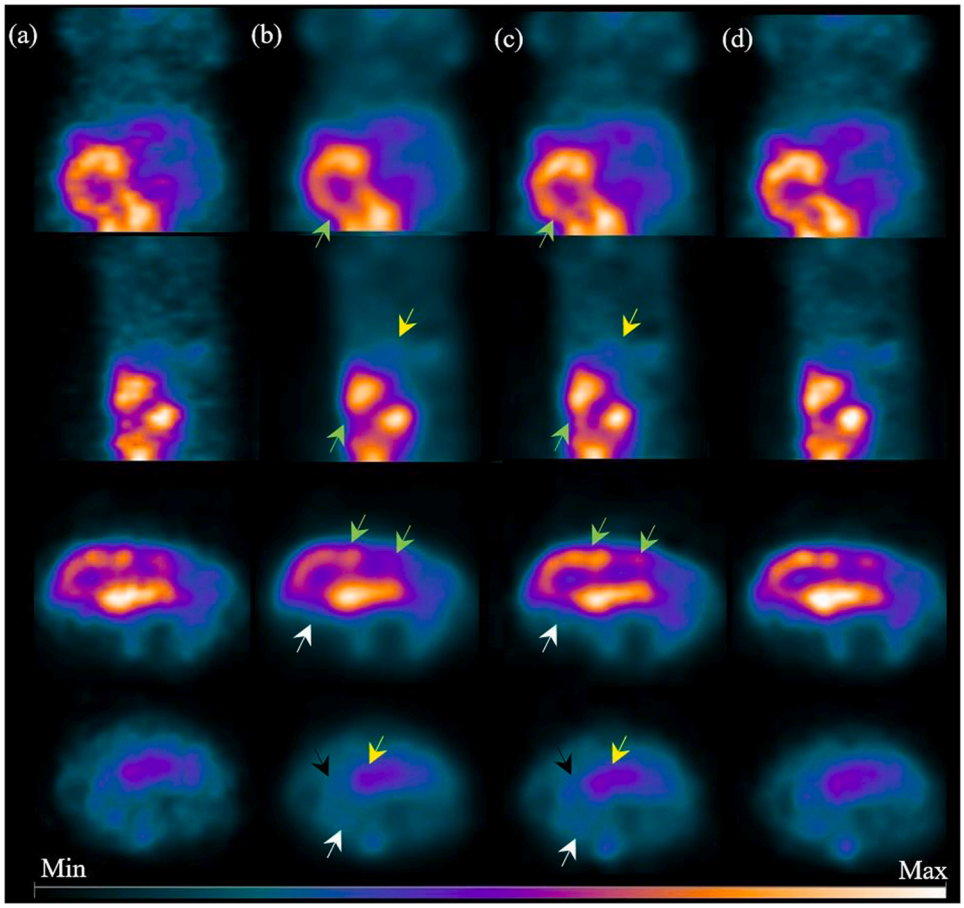

Fig. 6. Comparison between U-Net and PTWGAN. From top to bottom: representative PET images of a mouse from coronal, sagittal, and the transverse views from abdomen and liver, respectively. From left to right: (a) noisy LD images, (b) predicted images of U-Net, (c) predicted images of PTWGAN, and (d) original FD image (ground-truth). White arrows: kidney, Yellow arrows: pancreas, Green arrows: intestine, Black arrows: cranial part of duodenum recovered by PTWGAN. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pharmacological violation caused by mass effect and also the tracer-induced toxicity in animals. The situation is of huge importance, for example, when conducting quantitative studies with lower specific activities, shorter half-life radioisotopes, screening novel imaging agents with lower radiochemical yield, or in applications needed short frame duration (e.g., dynamic imaging and tracer kinetic studies). Also important is that injecting only small amounts of activity per scan could further improve the final image quality. This is attributed to a negligible fraction of random coincidences as well as the reduced dead-time effect in low activity regimens or through reducing the motion-related artifacts in rapid imaging scenarios. From an economic standpoint, moreover, low activity PET imaging is deemed cost-saving as it limits the amount of tracer per subject yet increases the animal throughput. So far, various dose minimization strategies were followed in the research environment. Molinos et al. reported on reducing tracer dose in animal studies by a factor of 4, using a total body microPET/CT scanner with extended axial FOV (150 mm) and improved detection sensitivity (Molinos et al., 2019). Liu et al. introduced a new post-reconstruction filtering approach and implemented it in nude mice studies (Liu et al., 2016). Hashimoto et al. proposed an unsupervised learning method for denoising the dynamic PET images collected from a living monkey brain and a series of simulation studies. Our results are in line with these findings and demonstrated that a 5-fold reduction in injected activity would be feasible using a DL-based framework even for a preclinical PET scanner with limited axial coverage ($\sim$ 50 mm). Towards this end, we trained five individual networks namely: 2D-UNet, 2D-RUNet, 2.5D-UNet, 2.5D-RUNet, and a PTWGAN framework with a 2.5D-UNet generator to convert the noise-corrupted LD images to regular quality FD mice scans. All networks were trained successfully and exhibited strong noise elimination capabilities without significant degradation in edges and corners. Among all U-Net models, the 2.5D network based on residual formulation surpassed other variants both numerically and perceptually. Using the same U-Net topology as the generator in a PTWGAN framework trained based on mixed loss function yielded a better approximation of tracer distribution and activity recovery. In contrast to PTWGAN, U-Net-based processing results in lower noise levels but slightly poor textural information.

Our findings also agree well with benchmark studies performed on human clinical PET datasets. Kaplan and Zhu (2019) applied a GAN framework for WB PET denoising and reported an average improvement of $\sim$ 30.98%, $\sim$ 2.02%, and $\sim$ 8.84% in RMSE, SSIM, and PSNR metrics, respectively. Our PTWGAN model also achieved 33.01%, 2.5%, and 4.81% improvement in terms of NRMSE, SSIM, and PSNR, respectively. Lei et al. (2019) obtained an average NCC of 0.975 ± 0.008 and PSNR of 39.3 ± 4 using a GAN model in WB PET denoising while the LD PET images had an average NCC of 0.963 ± 0.003 and PSNR of 38.1 ± 3.4. In the current study, we calculated an average PSNR of 36.89 ± 0.6 and NCC of 0.987 ± 0.0007 for PTWGAN outputs while the respective values were 35.2 ± 0.6 and 0.96 ± 0.0025 for our LD dataset.

As compared to already established filtering techniques, DL-based data recovery allowed us to achieve better results by reducing NRMSE values and improving PSNR, NCC, and SSIM measures. Considering varying noise levels among subjects, we tuned the filtering parameters to find the optimal results, which is often perplexing, time-consuming, and more importantly, requires domain expertise. For the case of BMNLM and ADF filters, we also eliminated the spurious signals through a median kernel to help the filtering process and to decrease the associated artifacts (Chan et al., 2014; Chan et al., 2007). On the contrary, DL models were trained on the grainy LD images and achieved satisfactory results without the need for computationally intensive processing or human assistance. These findings are in line with previous reports bore out the superiority of DL methods relative to existing post-denoising

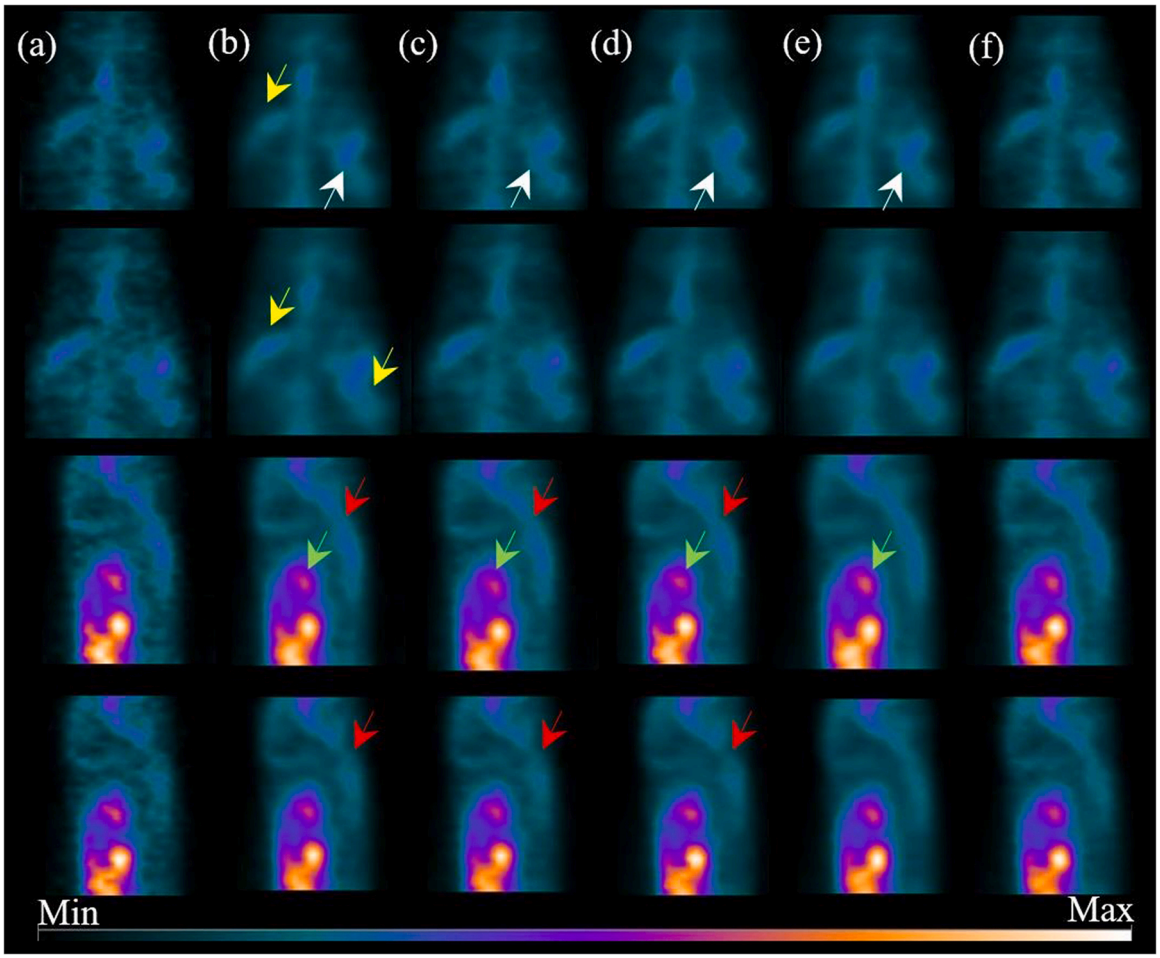

Fig. 7. Comparison among different denoising methods. Representative PET images of a mouse from coronal and sagittal views. From left to right: (a) noisy LD images, (b–d) denoised images with GF, ADF, and BMNLM, respectively, (e) DL (2.5D-RUNet) prediction, (f) original FD image (ground-truth). Red arrows: para-spinal region and the related artifacts, White arrows: uterine horn, Yellow arrows: blurred boundaries caused by GF around the lateral lobe of the liver and uterine horn of the kidney, Green arrows: underestimations in gastrointestinal tract caused by traditional filtering which is recovered by DL methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

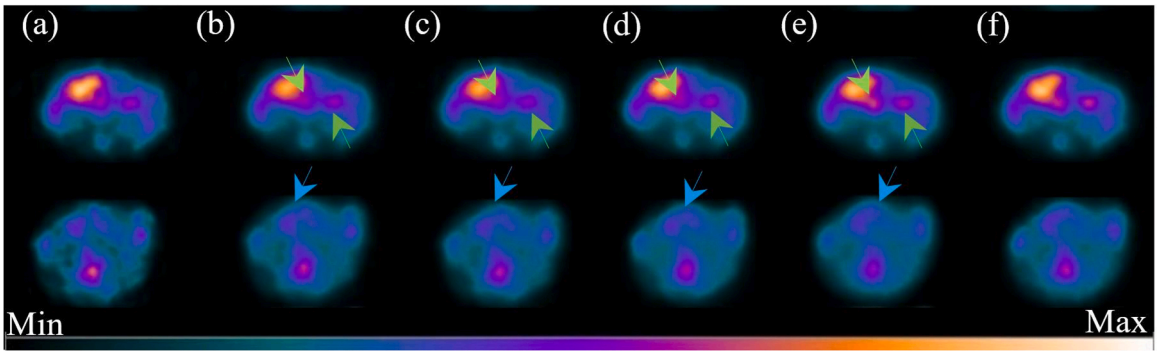

Fig. 8. Comparison among different denoising methods. Representative PET images of a mouse from abdominal (top) and chest (bottom) sections. From left to right: (a) noisy LD images, (b–d) denoised images with GF, ADF, and BMNLM, respectively, (e) DL (2.5D-RUNet) , (f) original FD image (ground-truth). Green arrows: gastrointestinal tract, Blue arrows: myocardium. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

methods (Gondara, 2016; Hashimoto et al., 2019; He et al., 2021; Lu et al., 2019; Xu et al., 2017).

Lu et al. (2019) investigated the quantitative accuracy of small lung nodules using DL-based denoising methods in oncological PET imaging. The authors showed that U-Net provides relatively smaller SUV bias as compared to Gaussian and NLM post-filtering. Hashimoto et al. (2019) demonstrated that noisy brain images treated by DL-based techniques have higher PSNR and SSIM relative to those processed with other algorithms, including Gaussian, IGF, and NLM filtering. Contrary to previous studies, where the model learns the mapping between low-count and high-count PET images, He et al. proposed a novel DL-based joint filtering framework, which learns the linear attenuation coefficients of

Table 3
Quantitative comparison among Different denoising methods. Quality metrics were calculated over all samples (fourteen mice) and reported as Mean- ± standard deviation.

	NRMSE	PSNR	NCC	SSIM
LD	0.115 ± 0.016	34.76 ± 2.1	0.965 ± 0.008	0.962 ± 0.007
DL	**0.073 ± 0.009**	**36.71 ± 2.01**	**0.986 ± 0.003**	**0.985 ± 0.003**
GF	0.084 ± 0.011	36.13 ± 2.04	0.98 ± 0.004	0.978 ± 0.004
ADF	0.0787 ± 0.009	36.41 ± 1.94	0.982 ± 0.003	0.982 ± 0.004
BMNLM	0.08 ± 0.008	36.28 ± 1.87	0.981 ± 0.004	0.981 ± 0.004
P-value				
LD vs. DL	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
LD vs. GF	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
LD vs. ADF	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
LD vs. BMNLM	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
DL vs. GF	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
DL vs. ADF	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
DL vs. BMNLM	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
GF vs. ADF	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
GF vs. BMNLM	< 0.02	< 0.02	0.137	**< 0.0001**
ADF vs. BMNLM	**< 0.0001**	**< 0.0001**	0.06	**< 0.01**

spatially variant linear representation model in dynamic PET image denoising (He et al., 2021). They reported that DL-based joint filtering provides more realistic estimates of standard-dose PET images and improves the results (RMSE = 0.138 ± 0.016, SSIM = 0.805 ± 0.034) over conventional CNN-only (RMSE = 0.15 ± 0.016, SSIM = 0.774 ± 0.037) and MRI-guided post-filtering (RMSE = 0.149 ± 0.014, SSIM = 0.741 ± 0.041) alone. Since the model uses the MRI information as a prior, a similar setup could be implemented to boost the performance of DL-based denoising in multimodal preclinical PET imaging. It would also be of great interest to compare our DL-based noise reduction method to other filtering approaches, such as wavelet (Ouahabi, 2013) and curvelet transforms (Bal et al., 2019) or jointly incorporating them within a DL framework to further enhance the network performance (Kang et al., 2018, 2017, 2021).

As reported in previous studies (Kang et al., 2017; Lu et al., 2019), the main culprit for DL-based image recovery seems to be the fact that the ground-truth images are also plagued by some noise, which might confound the learning process and leads to poor predictions. One could facilitate the training process by suppressing the noise both in source and target images via a low-passing filter or using a pre-trained network as implemented in this paper.

Further enhancement in the results might be possible, for instance, through jointly incorporating the anatomical outlines and structural information from other examination modalities (if available) as additional input channels, enlarging the sample size, using dilated convolutions (Ouahabi and Taleb-Ahmed, 2021; Spuhler et al., 2019), loss function optimization methods (Ouahabi and Taleb-Ahmed, 2021), retraining the pre-trained network using the LD images taken at different noise levels to cope with various noise intensity situations in PET studies (Kang et al., 2017). More sophisticated frameworks (e.g., iterative deep learning networks, direct reconstruction networks, etc.) or other GAN variants such as Cycle GAN may help to take major steps forward in performance and are worth exploring along this direction (Zhou et al., 2020).

Moreover, due to its miniaturized size, each tissue in mice spans just a limited number of slices. This made us train the networks using all the slices encompassing the trunk and upper abdomen as it is impossible to design more specific models for each organ. When compared to clinical practices, strict ethical issues, rapid metabolism process, physiological motion, and relatively larger variations in tracer uptake patterns among animal models add even more to the complexity and scarcity of the sample size in the context of preclinical imaging.

It is worth highlighting that we examined different setups and parameters to achieve the best scenario and optimum settings in our study. Since simple auto-encoders fail to converge at higher noise levels (Gondara, 2016), we faced several challenges when using basic CNNs and conventional U-Net architectures, including the appearance of severe checkboard artifacts, non-uniformities across the transaxial slices as well as resolution loss associated with standard CNNs. To mitigate these issues, we brought a number of modifications to the model to

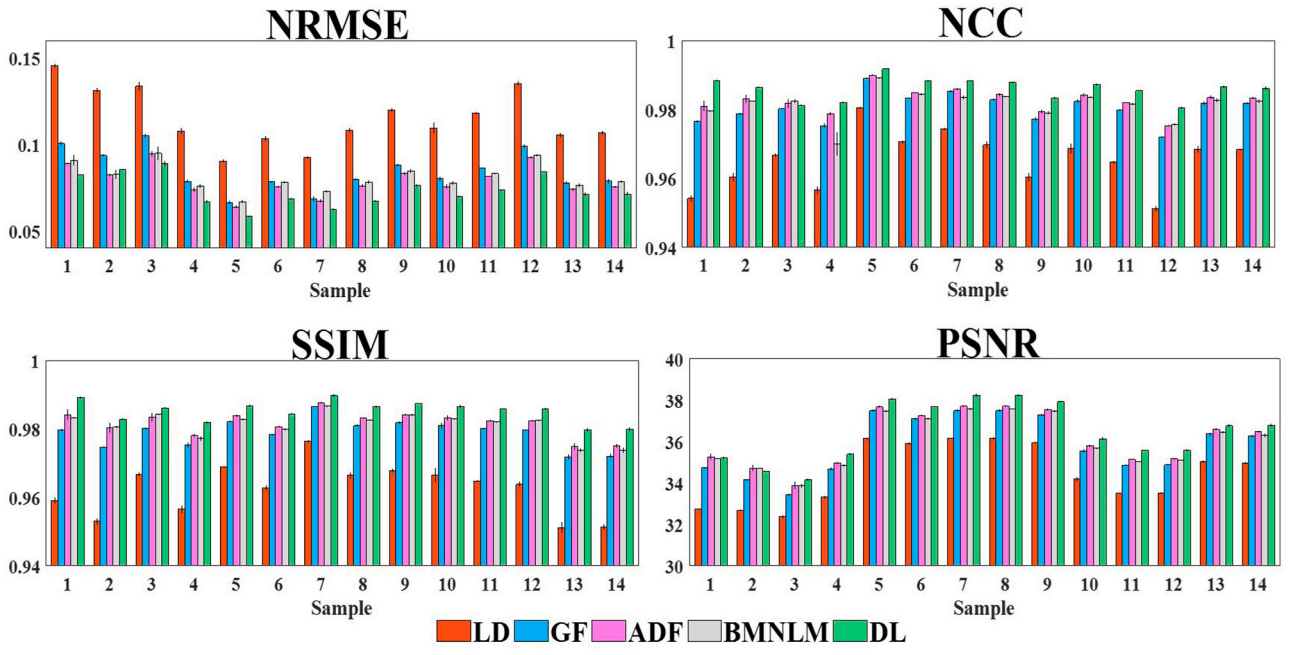

Fig. 9. Quantitative comparison among different denoising methods in the leave-one-group-out cross-validation study. From left to right and top to bottom: NRMSE, NCC, SSIM, and PSNR. For each subject, values averaged across five realizations. Error bars indicate standard deviation.

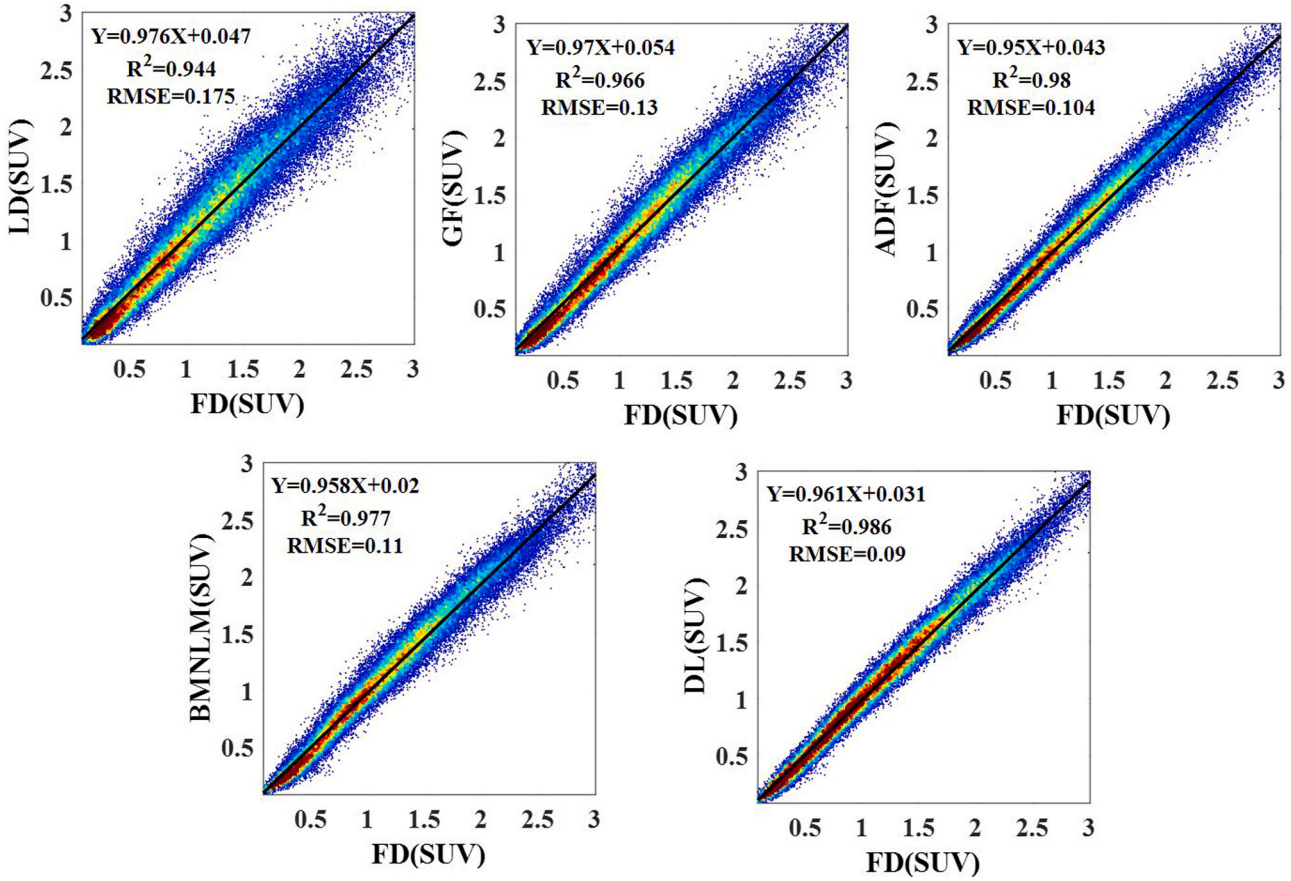

Fig. 10. Joint histogram analysis of SUV for different denoising strategies. From up to bottom and left to right: scatter and regression line plots for LD, GF, ADF, BMNLM, and DL (2.5D-RUNet) versus corresponding reference images.

obtain a simple yet effective design for our target task.

Recently, numerous DL-based strategies were established and implemented for noise reduction in different fields of biomedical imaging (Ouahabi and Taleb-Ahmed, 2021; Wang et al., 2021). However, choosing the best model for a specific task relies on different factors (e.g., number of training samples, corruption level or quality of the input dataset, number of input channels, availability of anatomical information, etc.). Other setups could be applied for preclinical PET image denoising and will be explored in future studies.

Given that the injected activity in small animal studies is much lower than in humans, herein, a dose reduction factor (DRF) of 80% was chosen to yield LD PET scans. We examined various ratios to determine the lowest achievable DRF. We found that, in our case (considering the initial injected dose, animal weight, imaging protocol, scanner performance, and, axial FOV), DRFs $< 80\%$ drastically alters the data composition in the LD mice images and leads to loss of texture information, which thwarts the learning process and impedes a meaningful mapping, particularly for the chest region.

Another important point worth mentioning is that we decimated the LD images by randomly down-sampling the events collected during a standard-dose acquisition. We have examined the reliability of the approach through several phantom experiments before conducting this study. By decreasing the activity concentration by a factor of 5–6 (real LD condition), we obtained 20–25% of the true rates when imaging the same phantom in high-dose conditions. Given that voxel-wise alignment between separate LD and FD scans is difficult, if not impossible to achieve (owing to motion and differences in tracer distribution between two separate acquisitions), almost all studies, except a few in the context of LD PET imaging, followed the same procedure to simulate low-dose conditions.

5. Conclusion

We investigated the feasibility of DL-enabled approaches to formulate standard-dose PET images from low-dose mice PET studies. A meaningful enhancement was found between LD images and those treated by different denoising techniques while the 2.5D DL models (trained with either MAE or mixed adversarial loss) performed the best amongst all investigated methods. However, more heterogeneous samples are needed to prove the reliability of our conjecture but our findings at least hint on that DL enables a promising direction toward low-count small animal studies without further need for expensive instrumentation or any error-prone post-processing steps. Compared to other conventional methods, DL brings a multitude of advantages not only in aspects of image quality, but also in terms of implementation, speed, and performance.

CRediT authorship contribution statement

Mahsa Amirrashedi: Conceptualization, Methodology, Data acquisition, Software, Investigation, Writing – original draft, Visualization. **Saeed Sarkar:** Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content. **Hojjat Mamizadeh:**

Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content. **Hossein Ghadiri:** Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content. **Pardis Ghafarian:** Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content. **Habib Zaidi:** Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content, Supervision, Funding acquisition. **Mohammad Reza Ay:** Conception and design of study, Analysis and/or interpretation of data, drafting the manuscript, revising the manuscript critically for important intellectual content, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grant No. 36847 from Tehran University of Medical Sciences, and the Swiss National Science Foundation under grant SNSF 320030_176052 and the Private Foundation of Geneva University Hospitals under Grant RC-06-01.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv 1603, 04467.

Amirrashedi, M., Ay, M.R., Sarkar, S., Farahani, M.H., 2020a. Normalization of a positron emission tomography scanner. US16/746,447. Parto Negar Persia (pnp) Co.

Amirrashedi, M., Sarkar, S., Ghafarian, P., Hashemi Shahraki, R., Geramifar, P., Zaidi, H., Ay, M.R., 2019. NEMA NU-4 2008 performance evaluation of Xtrim-PET: a prototype SiPM-based preclinical scanner. Med. Phys. 46, 4816–4825.

Amirrashedi, M., Zaidi, H., Ay, M.R., 2020b. Advances in preclinical PET instrumentation. PET Clin. 15, 403–426.

Amirrashedi, M., Zaidi, H., Ay, M.R., 2020c. Towards quantitative small-animal imaging on hybrid PET/CT and PET/MRI systems. Clin. Transl. Imaging 8, 1–21.

Arabi, H., Zaidi, H., 2018. Improvement of image quality in PET using post-reconstruction hybrid spatial-frequency domain filtering. Phys. Med. Biol. 63, 215010 https://doi.org/10.1088/1361-6560/aae573.

Arabi, H., Zaidi, H., 2020. Spatially guided nonlocal mean approach for denoising of PET images. Med. Phys. 47, 1656–1669.

Arabi, H., AkhavanAllaf, A., Sanaat, A., Shiri, I., Zaidi, H., 2021. The promise of artificial intelligence and deep learning in PET and SPECT imaging. Phys. Med. 83, 122–137.

Bal, A., Banerjee, M., Sharma, P., Maitra, M., 2019. An efficient wavelet and curvelet-based PET image denoising technique. Med. Biol. Eng. Comput. 57, 2567–2598.

Buades, A., Coll, B., Morel, J.-M., 2005. A non-local algorithm for image denoising. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2, pp. 60–65.

Chan, C., Fulton, R., Cai, W., Feng, D.D., Meikle, S., 2007. Minimum cross-entropy reconstruction of PET images with anatomically based anisotropic median-diffusion filtering. In: Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6527–6530.

Chan, C., Fulton, R., Barnett, R., Feng, D.D., Meikle, S., 2014. Postreconstruction nonlocal means filtering of whole-body PET with an anatomical prior. IEEE Transl. Med. Imaging 33, 636–650.

Chen, K.T., Gong, E., de Carvalho Macruz, F.B., Xu, J., Boumis, A., Khalighi, M., Poston, K.L., Sha, S.J., Greicius, M.D., Mormino, E., 2019. Ultra–low-dose 18F-florbetaben amyloid PET imaging using deep learning with multi-contrast MRI inputs. Radiology 290, 649–656.

Chollet, F., 2018. Keras: The Python Deep Learning Library. ASCL.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv 1511, 07289.

Cui, J., Gong, K., Guo, N., Wu, C., Meng, X., Kim, K., Zheng, K., Wu, Z., Fu, L., Xu, B., 2019. PET image denoising using unsupervised deep learning. Eur. J. Nucl. Med. Mol. Imaging 46, 2780–2789.

da Costa-Luis, C.O., Reader, A.J., 2021. Micro-networks for robust MR-guided low count PET imaging. IEEE Transl. Radiat. Plasma Med. Sci. 5, 202–212.

Gondara, L., 2016. Medical image denoising using convolutional denoising autoencoders. In: Proceedings of the 2016 IEEE 16th international conference on data mining workshops (ICDMW), 94, pp. 241–246.

Gong, K., Berg, E., Cherry, S.R., Qi, J., 2020a. Machine learning in PET: from photon detection to quantitative image reconstruction. Proc. IEEE 108, 51–68.

Gong, Y., Shan, H., Teng, Y., Tu, N., Li, M., Liang, G., Wang, G., Wang, S., 2020b. Parameter-transferred wasserstein generative adversarial network (PT-WGAN) for low-dose PET image denoising. IEEE Transl. Radiat. Plasma Med. Sci. 5, 213–223.

Han, Y.S., Yoo, J., Ye, J.C., 2016. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis. arXiv preprint arXiv 1611, 06391.

Hashimoto, F., Ohba, H., Ote, K., Teramoto, A., Tsukada, H., 2019. Dynamic PET image denoising using deep convolutional neural networks without prior training datasets. IEEE Access 7, 96594–96603.

He, Y., Cao, S., Zhang, H., Sun, H., Wang, F., Zhu, H., Lv, W., Lu, L., 2021. Dynamic PET image denoising with deep learning-based joint filtering. IEEE Access 9, 41998–42012. https://doi.org/10.1109/ACCESS.2021.3064926.

He, K., Sun, J., Tang, X., 2012. Guided image filtering. IEEE Trans. Pattern. Anal. Mach. Intell. 35, 1397–1409.

Herfert, K., Mannheim, J.G., Kuebler, L., Marciano, S., Amend, M., Parl, C., Napieczynska, H., Maier, F.M., Vega, S.C., Pichler, B.J., 2020. Quantitative rodent brain receptor imaging. Mol. Imaging Biol. 22 (2), 223–244.

Jagoda, E.M., Vaquero, J.J., Seidel, J., Green, M.V., Eckelman, W.C., 2004. Experiment assessment of mass effects in the rat: implications for small animal PET imaging. Nucl. Med. Biol. 31 (6), 771–779.

Kang, E., Min, J., Ye, J.C., 2017. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med. Phys. 44 (10), e360–e375. https://doi.org/10.1002/mp.12344.

Kang, E., Chang, W., Yoo, J., Ye, J.C., 2018. Deep convolutional framelet denosing for low-dose CT via wavelet residual network. IEEE Transl. Med. Imaging 37, 1358–1369. https://doi.org/10.1109/TMI.2018.2823756.

Kang, S.-K., Yie, S.-Y., Lee, J.-S., 2021. Noise2Noise improved by trainable wavelet coefficients for PET denoising. Electronics 10 (13), 1529.

Kaplan, S., Zhu, Y.-M., 2019. Full-dose PET image estimation from low-dose PET image using deep learning: a pilot study. J. Digit. Imaging 32, 773–778.

Kung, M.P., Kung, H.F., 2005. Mass effect of injected dose in small rodent imaging by SPECT and PET. Nucl. Med. Biol. 32, 673–678.

Lee, J.S., 2021. A review of deep-learning-based approaches for attenuation correction in positron emission tomography. IEEE Transl. Radiat. Plasma Med. Sci. 5, 160–184.

Lei, Y., Dong, X., Wang, T., Higgins, K., Liu, T., Curran, W.J., Mao, H., Nye, J.A., Yang, X., 2019. Whole-body PET estimation from low count statistics using cycle-consistent generative adversarial networks. Phys. Med. Biol. 64, 215017.

Liu, H., Wang, K., Tian, J., 2016. Postreconstruction filtering of 3D PET images by using weighted higher-order singular value decomposition. Biomed. Eng. Online 15, 102.

Lu, W., Onofrey, J.A., Lu, Y., Shi, L., Ma, T., Liu, Y., Liu, C., 2019. An investigation of quantitative accuracy for deep learning based denoising in oncological PET. Phys. Med. Biol. 64, 165019.

Molinos, C., Sasser, T., Salmon, P., Gsell, W., Viertl, D., Massey, J.C., Minczuk, K., Li, J., Kundu, B.K., Berr, S., Correcher, C., Bahadur, A., Attarwala, A.A., Stark, S., Junge, S., Himmelreich, U., Prior, J.O., Laperre, K., Van Wyk, S., Heidenreich, M., 2019. Low-dose imaging in a new preclinical total-body PET/CT scanner. Front. Med. 6, 88.

Ouahabi, A., Taleb-Ahmed, A., 2021. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. Pattern Recognit. Lett. 144, 27–34. https://doi.org/10.1016/j.patrec.2021.01.010.

Ouahabi, A., 2013. A review of wavelet denoising in medical imaging. In: Proceedings of the 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), pp. 19–26.

Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Transl. Pattern Anal. Mach. Intell. 12 (7), 629–639.

Reader, A.J., Zaidi, H., 2007. Advances in PET Image Reconstruction. PET Clin. 2, 173–190.

Reader, A.J., Corda, G., Mehranian, A., Costa-Luis, Cd, Ellis, S., Schnabel, J.A., 2021. Deep learning for PET image reconstruction. IEEE Transl. Radiat. Plasma Med. Sci. 5, 1–25.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical image computing and computer-assisted intervention, pp. 234–241.

Sanaat, A., Arabi, H., Mainta, I., Garibotto, V., Zaidi, H., 2020. Projection space implementation of deep learning-guided low-dose brain PET imaging improves performance over implementation in image space. J. Nucl. Med. 61, 1388–1396.

Sanaat, A., Shiri, I., Arabi, H., Mainta, I., Nkoulou, R., Zaidi, H., 2021. Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging. Eur. J. Nucl. Med. Mol. Imaging 48, 2405–2415.

Schaefferkoetter, J., Nai, Y.H., Reilhac, A., Townsend, D.W., Eriksson, L., Conti, M., 2019. Low dose positron emission tomography emulation from decimated high statistics: a clinical validation study. Med. Phys. 46, 2638–2645.

Serrano-Sosa, M., Spuhler, K., DeLorenzo, C., Huang, C., 2020. PET image denoising using structural MRI with a novel dilated convolutional neural network. J. Nucl. Med. 61, 434.

Shan, H., Zhang, Y., Yang, Q., Kruger, U., Kalra, M.K., Sun, L., Cong, W., Wang, G., 2018. 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. IEEE Transl. Med. Imaging 37, 1522–1534.

Spuhler, K., Serrano-Sosa, M., Cattell, R., DeLorenzo, C., Huang, C., 2019. Full-count PET recovery from low-count image using a dilated convolutional neural network. arXiv preprint arXiv 1910, 11865.

Tomasi, C., Manduchi, R., 1998. Bilateral filtering for gray and color images. In: Proceedings of Sixth international conference on computer vision (IEEE Cat. No. 98CH36271), pp. 839–846.

Wang, T., Lei, Y., Fu, Y., Wynne, J.F., Curran, W.J., Liu, T., Yang, X., 2021. A review on medical imaging synthesis using deep learning and its clinical applications. J. Appl. Clin. Med. Phys. 22, 11–36.

Wang, Y., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., Shen, D., Zhou, L., 2018. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. NeuroImage 174, 550–562.

Xiang, L., Qiao, Y., Nie, D., An, L., Wang, Q., Shen, D., 2017. Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose PET/MRI. Neurocomputing 267, 406–416.

Xu, J., Gong, E., Pauly, J., Zaharchuk, G., 2017. 200x low-dose PET reconstruction using deep learning. arXiv preprint arXiv 1712, 04119.

Zaidi, H., El Naqa, I., 2021. Quantitative molecular positron emission tomography imaging using advanced deep learning techniques. Annu. Rev. Biomed. Eng. 23, 249–276.

Zhou, L., Schaefferkoetter, J.D., Tham, I.W.K., Huang, G., Yan, J., 2020. Supervised learning with cyclegan for low-dose FDG PET image denoising. Med. Image Anal. 65, 101770.