Journal of Digital Imaging (2022) 35:469–481
https://doi.org/10.1007/s10278-021-00536-0

Robust-Deep: A Method for Increasing Brain Imaging Datasets to Improve Deep Learning Models' Performance and Robustness

Amirhossein Sanaat[1] · Isaac Shiri[1] · Sohrab Ferdowsi[2] · Hossein Arabi[1] · Habib Zaidi[1,3,4,5]

Received: 22 June 2021 / Revised: 29 September 2021 / Accepted: 8 November 2021 / Published online: 8 February 2022

Abstract

A small dataset commonly affects generalization, robustness, and overall performance of deep neural networks (DNNs) in medical imaging research. Since gathering large clinical databases is always difficult, we proposed an analytical method for producing a large realistic/diverse dataset. Clinical brain PET/CT/MR images including full-dose (FD), low-dose (LD) corresponding to only 5 % of events acquired in the FD scan, non-attenuated correction (NAC) and CT-based measured attenuation correction (MAC) PET images, CT images and T1 and T2 MR sequences of 35 patients were included. All images were registered to the Montreal Neurological Institute (MNI) template. Laplacian blending was used to make a natural presentation using information in the frequency domain of images from two separate patients, as well as the blending mask. This classical technique from the computer vision and image processing communities is still widely used and unlike modern DNNs, does not require the availability of training data. A modified ResNet DNN was implemented to evaluate four image-to-image translation tasks, including LD to FD, LD+MR to FD, NAC to MAC, and MRI to CT, with and without using the synthesized images. Quantitative analysis using established metrics, including the peak signal-to-noise ratio (PSNR), structural similarity index metric (SSIM), and joint histogram analysis was performed for quantitative evaluation. The quantitative comparison between the registered small dataset containing 35 patients and the large dataset containing 350 synthesized plus 35 real dataset demonstrated improvement of the RMSE and SSIM by 29% and 8% for LD to FD, 40% and 7% for LD+MRI to FD, 16% and 8% for NAC to MAC, and 24% and 11% for MRI to CT mapping task, respectively. The qualitative/quantitative analysis demonstrated that the proposed model improved the performance of all four DNN models through producing images of higher quality and lower quantitative bias and variance compared to reference images.

Keywords Brain PET · Deep learning · Data augmentation · Low-dose · Attenuation correction

Introduction

The growth of deep neural network (DNN) applications in medical image analysis during the last decade, though remarkable, still faces a number of challenges, including the issue of small size datasets and the limited size of annotated samples (or reference data). These problems are crucial for supervised learning models and for tasks requiring paired images (e.g., prediction of full-dose (FD) from low-dose (LD) PET images [1–4, 42, 43], prediction of attenuation-corrected (MAC) from non-attenuation-corrected (NAC) PET images [5–7], MRI to pseudo-CT mapping [8, 9], etc.), along with models requiring labeled data (e.g., segmentation, prognosis, etc.) [10–12]. Numerous constraints both ethical and logistical challenge gathering large clinical databases and even in the absence of these issues, the process is complex and time-consuming.

✉ Habib Zaidi
habib.zaidi@hcuge.ch

1 Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland

2 University of Applied Sciences and Arts of Western, Geneva, Switzerland

3 Geneva University Neurocenter, Geneva University, 1205 Geneva, Switzerland

4 Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

5 Department of Nuclear Medicine, University of Southern Denmark, DK-500 Odense, Denmark

The size of the dataset used for training of a DNN model has a direct influence on the generalizability, robustness, and qualitative/quantitative performance of the model on the test data (unseen data). The generalizability of a DNN model dictates the magnitude of the difference observed in the performance of a model when evaluated on the training dataset versus unseen test dataset [13]. The model robustness refers to a property of how a DNN algorithm performs on a new independent (but similar) dataset. In other words, a robust algorithm exhibits similar errors on the training and test datasets [14, 15].

It is often claimed that larger datasets lead to a better DNN model [16]. To build a generalizable robust DNN model, it is necessary that the validation loss decreases with the training loss continuously. To deal with the small dataset problem, data augmentation (DA) is commonly used as an easy and practical strategy enabling to significantly increase the size of the dataset and avoid overfitting. The augmented data are simply different representations of existing data that can be produced by either data warping or oversampling. Data warping augmentation change/transform the existing data in such a way that their underlying information and the corresponding labels are preserved. Conversely, oversampling augmentation approaches synthesize completely new data from the existing training dataset. Merging/mixing images [17, 18], feature space augmentations [19], and generative adversarial networks (GANs) [18, 20, 21] are well-known strategies.

Although most previous studies and proposed DA methods were limited to non-clinical data, few studies used the same approaches for medical applications. Han et al. proposed a two-step GAN DA model for synthesizing brain MR images with/without tumors separately [21]. Hao and Ogawara also used a conditional GAN-based DA method to generate synthetic fundus images [22]. Other studies presented a deep convolutional GAN for disease classification (infiltration, atelectasis, and normal) from chest x-ray images [23]. Frid-Adar et al. developed a GAN-based DA to improve the performance of the proposed CNN for liver lesion classification [20]. More recently, a technique, referred to as CovidGAN, that exploits an auxiliary classifier GAN and synthetic data augmentation of training dataset was proposed to enhance Covid-19 detection [24].

In all the abovementioned studies, the GAN network was used to synthesize artificial data, mostly for classification tasks. If the dataset is not large enough, the training of the GAN models might not produce realistic synthetic images. More importantly, the synthesized images from imperfectly trained deep learning model would result in high quantitative bias.

To address these concerns, we proposed a novel technique, which relies on simple analytical methods that can produce realistic/diverse synthetic images from a limited dataset without using DNN models. A single study in the field of medical imaging has addressed this issue for bone segmentation from whole-body CT images [17]. The authors proposed an analytical DA technique to synthesize new images from four randomly selected images of a randomly cropped and patched dataset. However, the produced images are not realistic since different irrelevant parts of the body would be concatenated together.

In the current study, we focused on improving the outcome of four important image analysis applications in multimodality brain imaging by increasing the size of the training dataset. Our proposed technique, referred to as Robust-Deep, increases the number of brain imaging datasets for the different imaging modalities investigated in this work. The model combines registered images of two different patients and uses the Laplacian blending (LB) technique and an empirical sampling mask to produce realistic images from original images for all modalities. Robust-Deep improves the robustness of deep learning models by decreasing the bias while enhancing the qualitative and quantitative accuracy of the model. To the best of our knowledge, this is the first study reporting on a data augmentation technique designed for multimodality brain imaging addressing four different tasks, including LD-PET to FD-PET, LD-PET + MRI to FD-PET, NAC-PET to MAC-PET, and MR to CT image translations.

Materials and Methods

The performance of the proposed DA technique for four deep learning-guided image analysis tasks, including LD-PET to FD-PET, LD-PET + MRI (T1 and T2-weighted) to FD-PET, NAC-PET to MAC-PET, and MR to CT image conversions.

PET/CT and MRI Data Acquisition

The patient population consisted of 45 patients (19 males and 26 females, 63 ± 9 years and 71 ± 13 years, respectively) presenting with cognitive symptoms of possible neurodegenerative disease who underwent brain ^{18}F-FDG PET/CT and MRI examinations collected between April 2017 and September 2019 at Geneva University Hospital. The detailed demographic information of the patients is summarized in Table 1. The study protocol was approved by the institution's ethics committee and all patients gave written informed content. PET/CT acquisitions were performed on a Biograph mCT scanner (Siemens Healthcare, Erlangen, Germany) about 35 min post-injection. A low-dose CT scan (120 kVp, 20 mAs) was performed for PET attenuation correction. The patients underwent a 20-min static brain PET scan after injection of 205 ± 10 MBq of ^{18}F-FDG. PET data were acquired in list-mode format and

Table 1 Demographics of patients included in this study protocol

	Training	Test
Number of datasets	35	10
Male/female	15/20	4/6
Age (mean ± SD)	63±9	71 ± 13
Weight (mean ± SD)	72 ± 15	66 ± 14
Indication/diagnosis	Cognitive symptoms of possible neurodegenerative etiology	
Available modality	PET, CT, MRI	

reconstructed using the e7 tool (an offline reconstruction toolkit provided by Siemens Healthcare) to produce FD PET images. Subsequently, a subset of PET data containing 5% of the total events was extracted randomly from the list-mode data to produce the LD images using a validated code [25]. Both FD and LD PET images were reconstructed into a $200 \times 200 \times 109$ image matrix ($2.03 \times 2.03 \times 2.2$ mm^3 voxel size) using an ordinary Poisson ordered subsets-expectation maximization (OP-OSEM) algorithm (5 iterations, 21 subsets) with point spread function (PSF) modelling. PET images underwent post-reconstruction Gaussian filtering with 2 mm FWHM similar to the clinical protocol. MRI data acquisition was carried out on a 3T MAGNETOM Skyra (Siemens Healthcare, Erlangen, Germany) with a 64-channel head coil. The MRI scans included a 3D T1-weighted magnetization prepared rapid gradient-echo, MP-RAGE (TE/TR/TI, 2.3 ms/1930 ms/970 ms, flip angle 8°; NEX = 1, voxel size $0.8 \times 0.8 \times 1$ mm^3) and a 3D T2-weighted (TE/TR, 386 ms/5000 ms, NEX = 1; voxel size $0.5 \times 0.5 \times 1$ mm^3).

Data Preparation

To preserve the quantitative information in the images, the reconstructed images were first converted to the corresponding unit (standardized uptake value (SUV) for PET images and Hounsfield units (HUs) for CT images) and were then divided by constant values to map the intensities within the range [0 – 1]. Subsequently, the images were cropped, and all FD PET images registered to a brain ^{18}F-FDG template defined into standard Montreal Neurological Institute (MNI) stereotactic space [26] using the 3D Slicer software. A rigid registration method with 6 degrees of freedom was employed. Since PET-CTAC images are registered to a common template, the same transformation matrices were applied to LD, CT, MAC, and NAC images. In the next step, we registered MR to CT images to produce a dataset of 45 patients with well aligned FD, LD, CT, MAC, NAC, T1, and T2 images registered to the template.

Image Synthesis with Laplacian Blending

Consider two images $\mathbf{X}_1$ and $\mathbf{X}_2$, that we blend using a binary mask $\mathbf{M}$, i.e., our aim is to construct a blended image $\mathbf{Y}$ containing the contents of $\mathbf{X}_1$ at locations where $\mathbf{M} = 1$, and otherwise containing the contents of $\mathbf{X}_2$ where $\mathbf{M} = 0$.

A naïve solution to this problem is to directly blend the pixel values of these images using the mask, i.e., $\mathrm{Y} = \mathrm{M} \odot \mathrm{X}_1 + (1 - \mathrm{M}) \odot \mathrm{X}_2$ where $\odot$ denotes element-wise multiplication.

This pixel-domain approach, however, provides poor visual quality, since the transition between the contents of the two images may be very sharp at the borders of the mask.

A better alternative is LB, which attempts to make the transitions more natural using information from the frequency domain of the different images, and the blending mask. In this work, we used only two different patients' images, but it is possible to increase the number of patients. This classical technique within the computer vision and image processing communities is still widely used and, unlike modern deep learning techniques, does not require the availability of training data [27].

LB is based on Gaussian and Laplacian pyramids, which are series of images with increasingly smaller sizes derived from an input image at the base level, hence forming a pyramidal shape.

Concretely, consider the down-sampling operator $h_{\downarrow 2}(\cdot)$, where we form an output image with half the height and width of the input image (hence with 4 times fewer pixels) by taking every other pixel of the input image and neglecting the in-between pixels. The direct application of this operator to an image, however, mixes the different frequency contents of the image and as such, the output may be significantly different from the input. To avoid this phenomenon, known as aliasing in signal processing, before applying $h_{\downarrow 2}(\cdot)$, the image is passed through a low-pass filter $h_f(\cdot)$ to cut the high-frequency content of the image, hence avoiding the aliasing effect. A practical choice for this filter is the Gaussian function, and hence repeatedly applying the filtering followed by down-sampling, i.e., $h_{\downarrow 2}(g_f(\cdot))$ produces a Gaussian pyramid. Hence, in a multi-level Gaussian pyramid, the image at level i is formed from the image at level $i-1$ as $\mathbf{X}^{[i]} = h_{\downarrow 2}(g_f(\mathbf{X}^{[i-1]}))$, where superscript i indicates the level, and $\mathbf{X}^{[0]} = \mathbf{X}$, i.e., the base of the pyramid is the original image itself. Eq. 1 demonstrates an L-level Gaussian pyramid from the input image $\mathbf{X}$:

$$\begin{aligned} \mathbf{X}^{[0]} &= \mathbf{X}, \\ \mathbf{X}^{[1]} &= h_{\downarrow 2}(g_f(\mathbf{X}^{[0]})), \\ &\vdots \\ \mathbf{X}^{[L]} &= h_{\downarrow 2}(g_f(\mathbf{X}^{[L-1]})), \end{aligned} \tag{1}$$

Parallel to the downsampling operation, the upsampling operation $h_{\uparrow 2}(\cdot)$ produces an output image with twice the width and height of the input image by putting a zero-valued pixel between every two neighboring pixels (both horizontally and vertically). Again to avoid aliasing, a Gaussian low-pass filter should be applied to the image, but this time after the upsampling, i.e., $g_f(h_{\uparrow 2}(\cdot))$.

The Laplacian pyramid at level i, $\overset{\sim}{\mathbf{X}}^{[i]}$ is built through upsampling the $i+1^{\text{th}}$ level of the Gaussian pyramid and subtracting it from its i^{th} level, i.e., $\overset{\sim}{\mathbf{X}}^{[i]} = \mathbf{X}^{[i]} - g_f(h_{\uparrow 2}(\mathbf{X}^{[i+1]}))$. While the Gaussian pyramid contains the low-frequency content of the image, the Laplacian pyramid contains its high-frequency content, like edges and corners. Eq. 2 demonstrates an L-level Laplacian pyramid for the input image $\mathbf{X}$:

$$\begin{aligned}
\overset{\sim}{\mathbf{X}}^{[L]} &= \mathbf{X}^{[L]},\\
\overset{\sim}{\mathbf{X}}^{[L-1]} &= \mathbf{X}^{[L-1]} - g_f(h_{\uparrow 2}(\mathbf{X}^{[L]})),\\
&\vdots\\
\overset{\sim}{\mathbf{X}}^{[1]} &= \mathbf{X}^{[1]} - g_f(h_{\uparrow 2}(\mathbf{X}^{[2]})),\\
\overset{\sim}{\mathbf{X}}^{[0]} &= \mathbf{X}^{[0]} - g_f(h_{\uparrow 2}(\mathbf{X}^{[1]})).
\end{aligned} \tag{2}$$

LB is based on the above discussed Gaussian and Laplacian pyramids. Consider again the input images $\mathbf{X}_1$ and $\mathbf{X}_2$ and the blending mask $\mathbf{M}$. We construct the Gaussian and Laplacian pyramids for the input images, as well as a Gaussian pyramid for the mask [28]. The process of LB then consists of constructing a blended pyramid from these two Laplacian pyramids using weights derived from the Gaussian pyramid of the mask. Equation 3 shows the procedure followed to construct the Laplacian blending:

$$\begin{aligned}
\overset{\sim}{\mathbf{Y}}^{[L]} &= \mathbf{M}^{[L]} \odot \overset{\sim}{\mathbf{X}}_1^{[L]} + (1 - \mathbf{M}^{[L]}) \odot \overset{\sim}{\mathbf{X}}_2^{[L]},\\
\overset{\sim}{\mathbf{Y}}^{[L-1]} &= \mathbf{M}^{[L-1]} \odot \overset{\sim}{\mathbf{X}}_1^{[L-1]} + (1 - \mathbf{M}^{[L-1]}) \odot \overset{\sim}{\mathbf{X}}_2^{[L-1]} - g_f(h_{\uparrow 2}(\mathbf{Y}^{[L]})),\\
&\vdots\\
\overset{\sim}{\mathbf{Y}}^{[0]} &= \mathbf{M}^{[0]} \odot \overset{\sim}{\mathbf{X}}_1^{[0]} + (1 - \mathbf{M}^{[0]}) \odot \overset{\sim}{\mathbf{X}}_2^{[0]} - g_f(h_{\uparrow 2}(\mathbf{Y}^{[1]})),
\end{aligned} \tag{3}$$

where the final blended image is the base of this pyramid, i.e. $\overset{\sim}{\mathbf{Y}}^{[0]}$.

In this work, we tried five different masks (Fig. 1b) but report only the results associated with masks 1 and 2 in Fig. 1. The masks' pattern can be completely arbitrary. The reason behind choosing these two masks as a sample was to show a simple (mask #1) and a complex mask (mask #2) approach for synthesizing data. It is worth emphasizing that the process of designing the mask could be adopted according to the task and the endpoints of the study to create synthetic datasets bearing/sharing specific structures/features/anatomies with the original data.

Deep Learning Model

The ResNet model [29] employed in this work contains 20 convolutional layers wherein various levels of feature extraction are applied to the input images or feature maps using the kernel dilation concept. For low-level features extraction (the first seven layers), a convolution kernel with a size of 3 × 3 × 1 voxels and zero dilatation was selected. For medium-level features extraction (the next seven layers), a dilation of two was used whereas a dilation factor of four was considered for high-level feature extraction (the last six-layer). In this network, the rectified linear unit (ReLU) was used for activation function and every two convolutional layers were connected by a residual link. The input image/images

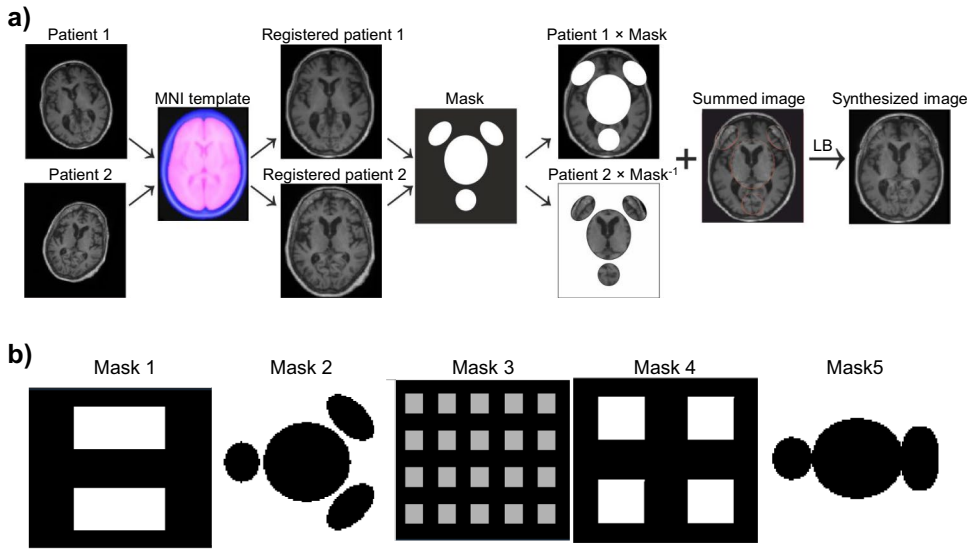

Fig. 1 **a** Schematic view of the proposed model for synthesizing realistic images through LB, and **b** the masks created to take samples from the input images. After a preliminary evaluation, masks #1 and #2 were selected for further investigation

(namely LD, LD+MRI, NAC, MRI) were fed to the ResNet network to predict target images (FD, FD, MAC, CT) in an end-to-end image translation fashion.

Our ResNet network consists of an encoder-decoder enabling image-to-image translation via dilated convolutional layers to map the inputs to the target images. The model involves three main sections consisting of a 3-level dilated convolution layers with the residual connection. The critical aspect of our modified ResNet, plus dilated convolution kernels, is the residual connections that bypass the parameterized layers. The ResNet model benefits from 9 residual blocks, which results in a large number of receptive fields that improved the feature extraction process.

The training for all image analysis tasks was performed using 35 patients with 5 fold cross-validation scheme, with each patient including LD, FD, NAC, MAC, CT, and MR (T1 and T2-weighted) images as input/output, respectively. The training of the ResNet model was performed using a 2D spatial window equal to 112×96 pixels and a batch size of 5. The following setting was used for the training of the two models with and without applying data augmentation: learning rate = 0.001, sample per volume = 1, optimizer = Adam, loss function = L2norm, and decay = 0.0001.

To generate synthetic CT images from T1- and T2-weighted MR sequences, a dual-channel ResNet model was developed to take the two MR images as input to predict the corresponding CT images. For consistency, the network's hyperparameters were kept similar to other networks. This enables to assess the effect of sample size independently from the network.

The model was implemented on NVIDIA 2080Ti GPU with 11 GB memory running under Windows 10 operating system. The training of the model for the four tasks was carried in 20 epochs. The training and hyperparameter fine-tuning of the model were performed on 35 patients (3'220 2D slices) and 5 fold cross-validation scheme. A separate unseen dataset consisting of 10 patients served as a test dataset (920 2D slices).

Quantitative Evaluation Strategy

The performance of our models was evaluated through estimation of the accuracy of the predicted images using three well-established quantitative metrics, including the root mean squared error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index metrics (SSIM) (Eqs. 3, 4 and 5, respectively).

$$RMSE(R,P) = \sqrt{\frac{\sum_{j=1}^{L}(R-P)^2}{L}} \tag{4}$$

$$PSNR(R,P) = 20 \times log_{10}^{\left(\frac{Max(P)}{\sqrt{MSE(R,P)}}\right)} \tag{5}$$

$$SSIM(R,P) = \frac{\left(2m_R m_P + c_1\right)(2\sigma_{RP} + c_2)}{(m_R^2 + m_P^2 + c_1)(\sigma_R^2 + \sigma_P^2 + c_2)} \tag{6}$$

In Eq. (3), L is the total number of voxels in the head region, R is the reference image (either FD, MAC, or CT), and P is the predicted image (either synthesized FD, MAC, or CT). In Eq. (4),$Max(P)$ indicates the maximum intensity value of R or P, whereas MSE is the mean squared error. m_R and m_P in Eq. (5) denote the mean value of images R and P, respectively. σ_{RP} indicates the covariance of σ_R and σ_P, which in turn represent the variances of R and P images, respectively. The constant parameters c_1 and c_2 ($c_1 = 0.01 \text{and} c_2 = 0.02$) were used to avoid a division by

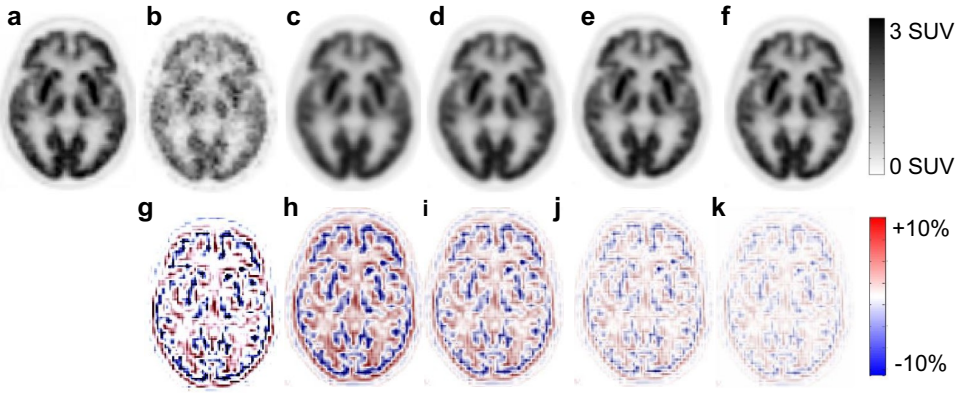

Fig. 2 The upper and lower panels show the reference and predicted FD PET images and the corresponding bias maps between the predicted images and the reference FD image, respectively. **a** Reference FD PET image, **b** LD PET image, and predicted images using limited (35 cases) **c** PU, and **d** PR. The predicted images belong to models fed by a large dataset (350 synthesized cases) **e** PM1 and **f** PM2. The bias maps for LD and PU, PR, PM1, and PM2 are shown in **g**–**k** images, respectively

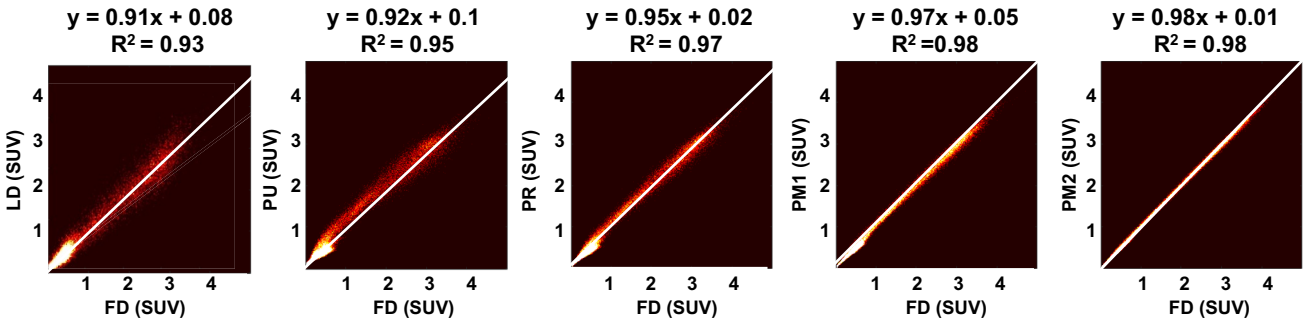

Fig. 3 Joint histogram analysis between reference FD and from left to right LD image and predicted FD images generated using PU, PR, PM1, and PM2 techniques, respectively

very small numbers. For LD to FD translation task, these metrics were also calculated for LD images to provide an insight into the noise levels and significant signals in the LD images.

Joint histogram analysis was also carried out to depict the voxel-wise correlation of the activity concentration between predicted and reference images. The statistical analysis was performed by calculating the pairwise Student's *t*-test for RMSE, SSIM, and PSNR between the different scenarios (predictions using mask #1 (PM1) vs. predictions using mask #2 (PM2) and predicted unregistered (PU) vs. predicted registered (PR)) using the MedCalc software [30]. The significance level was set at p value < 0.05 for all comparisons. PM1 and PM2 represent the predicted images by networks trained with two datasets including 35 real images plus 350 synthetic images (based on LB) using masks #1 and #2, respectively (Fig. 1). PU and PR are the predicted images using models trained with a limited dataset consisting of only the actual 35 images.

Results

Low-Dose to Full-Dose PET Image Translation

The synthesized images FD from LD images exhibited notable enhancement in image quality compared to LD images, providing almost similar appearance with respect to the reference FD PET images (Fig. 2). The visual inspection revealed that the predicted images derived from training with unregistered images (PU) are slightly blurred compared to those derived from the model fed by images registered to the MNI template (PR), although they still show great improvement compared to LD PET images. Though the two predicted images (PM1 and PM2) showed approximately similar image quality, PM2 images bear less quantitative bias compared to the other predicted images.

Figure 3 illustrates linear regression plots depicting the correlation between tracer uptake for LD, PU, PR, PM1, and PM2 with respect to FD. The scatter and linear regression plots showed higher correlation between PR and FD ($R^2 = 0.97$, slope = 0.95) compared to PU ($R^2 = 0.95$, slope = 0.92). Among the predicted images, PM2 achieved overall the best performance ($R^2 = 0.98$, slope = 0.98) while a poor correlation ($R^2 = 0.93$, slope=0.91) was observed for LD PET images.

Table 2 summarizes the RMSE, SSIM and PSNR calculated on the test dataset for LD, PU, PR, PM1, and PM2 PET images. Overall, the PR showed improved image quality and better noise properties with statistically significant differences compared PU.

Low-Dose PET + MRI to Full-Dose PET Image Translation

The use of MRI beside LD images enables to predict FD PET images with higher quality and more realistic anatomical information compared to the model relying only on LD

Table 2 Comparison of results obtained from the analysis of image quality in LD PET images and predicted FD PET images (PU, PR, PM1, PM2) from only LD PET images for the test dataset. SSIM: structural similarity index metrics, PSNR: peak signal to noise ratio, RMSE: root mean squared error

LD to FD	RMSE	SSIM	PSNR
LD	0.42 ± 3.24	0.62 ± 0.18	14.11 ± 4.32
Unregistered predictions (PU)	0.35 ± 0.12	0.82 ± 0.07	19.13 ± 1.32
Registered predictions (PR)	0.28 ± 0.08	0.86 ± 0.04	28.57 ± 1.55
Registered mask#1 (PM1)	0.23 ± 0.06	0.90 ± 0.02	34.07 ± 0.98
Registered mask#2 (PM2)	0.20 ± 0.05	0.93 ± 0.03	37.14 ± 0.10
P value (PM2 vs. PM1)	< 0.01	< 0.02	< 0.01
P value (PU vs. PR)	< 0.02	< 0.02	< 0.02

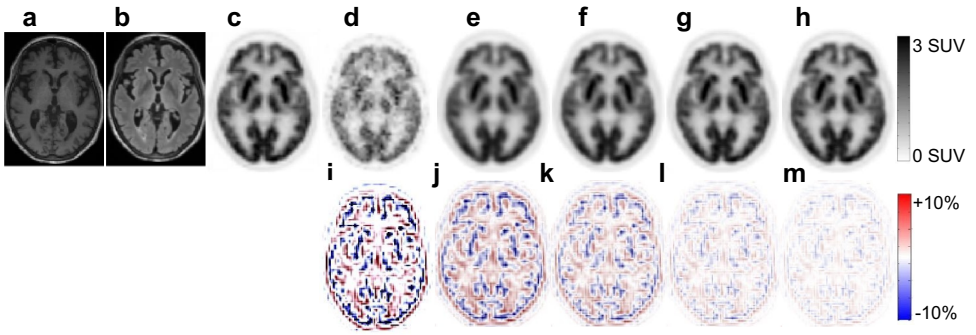

Fig. 4 The upper and lower panels show the reference and predicted FD PET images and the corresponding bias maps between the predicted images and the reference FD image, respectively. **a** T1-weighted MRI, **b** T2-weighted MRI, **c** reference FD PET image, **d** LD PET image, and predicted images using limited dataset (35 cases) **e** PU, and **f** PR. The predicted images belong to models fed by a large dataset (350 synthesized cases) **g** PM1 and **h** PM2. The bias maps for LD and PU, PR, PM1, and PM2 are shown in **i**–**m** images, respectively

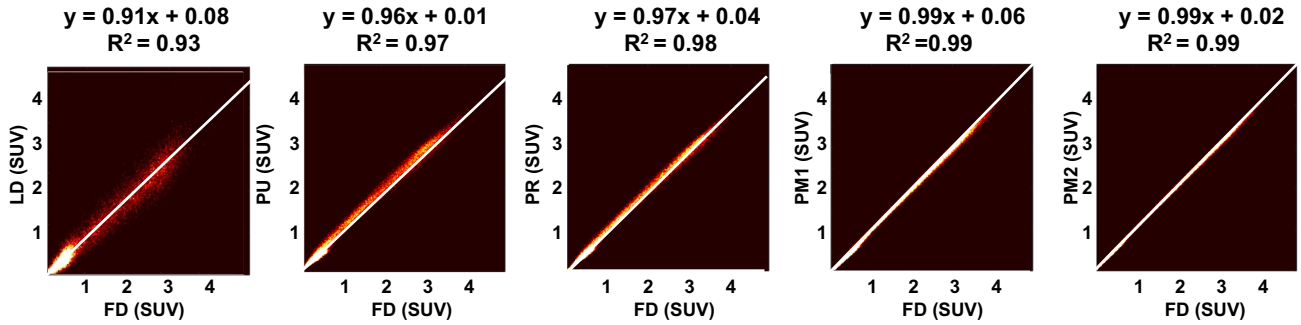

Fig. 5 Joint histogram analysis between reference FD and from left to right LD image and predicted FD images supported by MRI generated using PU, PR, PM1, and PM2 techniques, respectively

images (Fig. 2). Figure 4 displays representative transverse views of FD, LD, and predicted PET images along with the T1- and T2-weighted MR images. The visual inspection revealed that the images synthesized from the training data set produced by Robust-Deep (PM1 and PM2) better reflected the underlying FDG uptake patterns and anatomy than those obtained from small data set (PU and PR). The bias maps show that an increased dataset using the proposed method reduced quantitative bias.

For this task, the linear regression analysis performed over the test dataset (Fig. 5) demonstrated a high correlation for PM2 ($R^2 = 0.99$, slope = 0.99). A lower correlation and underestimation of tracer uptake was obtained ($R^2 = 0.97$, slope = 0.96) when utilizing unregistered datasets (PU) without using the proposed augmentation technique. Moreover, the SSIM, PSNR, and RMSE metrics calculated between LD and predicted images versus FD images are reported in Table 3. The quantitative evaluation metrics demonstrated the superior performance of the proposed data augmentation technique which supplied the deep learning model with realistic synthetic images in addition to the original training dataset.

Non-attenuation-Corrected to Attenuation-Corrected PET Image Conversion

Figure 6 presents reference MAC, NAC, and predicted MAC PET images produced from NAC using small unregistered and registered dataset and large dataset using mask number

Table 3 Comparison of results obtained from the analysis of image quality in LD PET images and predicted FD PET images (PU, PR, PM1, PM2) from LD PET + MR images for the test dataset. SSIM: structural similarity index metrics, PSNR: peak signal to noise ratio, RMSE: root mean squared error

LD+MRI to FD	RMSE	SSIM	PSNR
LD	0.42 ± 3.24	0.62 ± 0.18	14.11 ± 4.32
Unregistered predictions (PU)	0.29 ± 0.08	0.85 ± 0.04	22.74 ± 2.98
Registered predictions (PR)	0.24 ± 0.06	0.90 ± 0.03	32.96 ± 2.31
Registered mask#1 (PM1)	0.15 ± 0.06	0.94 ± 0.03	38.83 ± 1.82
Registered mask#2 (PM2)	0.14 ± 0.03	0.97 ± 0.01	39.46 ± 0.68
P value (PM2 vs. PM1)	< 0.05	< 0.05	< 0.05
P value (PU vs. PR)	< 0.01	< 0.01	< 0.01

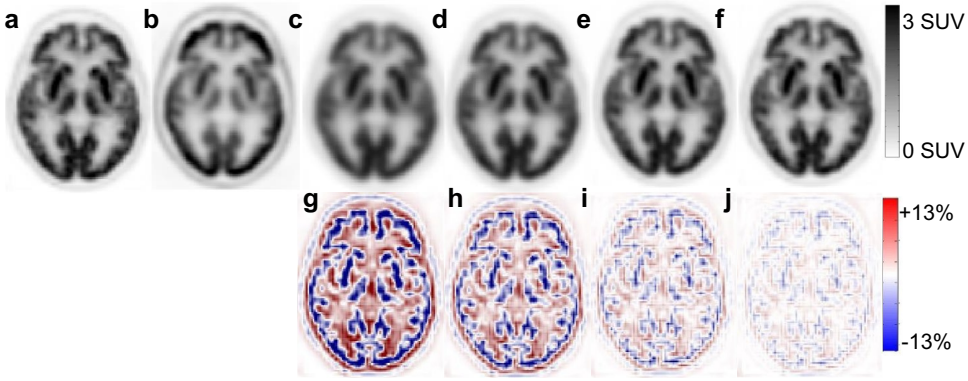

Fig. 6 The upper panel shows the reference and predicted images as well as bias maps. **a** Reference MAC PET, **b** NAC PET, and predicted AC PET images using **c** PU, **d** PR, **e** PM1, and **f** PM2. The bias maps for PU, PR, PM1, and PM2 are illustrated in **g**–**j**, respectively

one and two. The bottom panel shows bias maps of predicted images to the reference MAC. As can be seen from Fig. 6, PM2 images show the lowest bias compared to CT-based AC reference MAC image. Figure 7 depicts the joint histogram analysis between predicted AC PET images versus MAC PET images for PU, PR, PM1, and PM2. The lowest and highest correlations were achieved by PU ($R^2 = 0.96$, slope = 0.94) and PM2 ($R^2 = 0.99$, slope = 0.98), respectively. The quantitative analysis results are summarised in Table 4. One can see that PM1 and PM2 resulted in significantly lowest values of RMSE and highest values of PSNR and SSIM compared to PU and PR (P value < 0.05).

MRI to CT Image Conversion

Representative samples of the resulting CT images using the different data augmentation techniques are depicted in Fig. 8 wherein T1, T2 MR sequences, and reference CT image along with the corresponding bias maps are also presented. Substantially lower CT value bias was observed in synthetic CT images generated by Robust-Deep model. The results of the joint histogram analysis between the resulting synthetic and original CT images are presented in Fig. 9. Dramatic improvement was observed between PU ($R^2 = 0.91$, slope = 0.78) and the PM2 ($R^2 = 0.99$, slope = 0.94) models, demonstrating the effectiveness of the proposed data augmentation technique. In agreement with the joint histogram analysis, the quantitative metrics calculated on the resulting synthetic CT images in Table 5 demonstrated the superior performance of the Robust-Deep model, which achieved a mean RMSE of 0.22 ± 0.08 compared to 0.34 ± 0.09 obtained using the PU model.

Overall Performance and Robustness

The qualitative metrics, including SSIM, PSNR, and RMSE calculated for PM2 and PR models after each epoch separately for LD to FD PET conversion task are shown in Fig. 10. The quantitative metrics were only calculated for epochs residing in the plateau of the training loss. The results indicated that the PM2 model when the network is trained with a large dataset (350 synthesized + 35 real patients), is noticeably more robust/accurate with less fluctuations compared to PR trained only with 35 real patients.

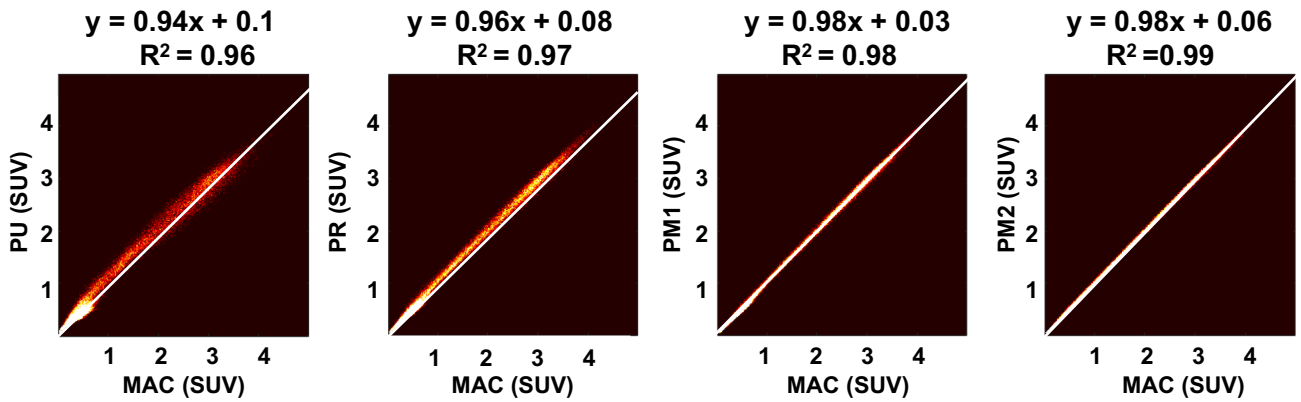

Fig. 7 Joint histogram analysis between reference MAC PET and from left to right predicted AC PET images generated using PU, PR, PM1, and PM2 techniques, respectively

Table 4 Comparison of results obtained from the analysis of image quality in predicted AC PET images (PU, PR, PM1, PM2) from non-attenuation corrected PET images for the test dataset. SSIM: structural similarity index metrics, PSNR: peak signal to noise ratio, RMSE: root mean squared error

NAC to MAC	RMSE	SSIM	PSNR
Unregistered (PU)	0.38 ± 0.13	0.82 ± 0.08	24.15 ± 3.82
Registered (PR)	0.32 ± 0.1	0.88 ± 0.08	26.57 ± 2.45
Registered mask#1 (PM1)	0.26 ± 0.08	0.94 ± 0.05	30.46 ± 3.01
Registered mask#2 (PM2)	0.26 ± 0.07	0.95 ± 0.06	31.08 ± 2.81
P value (PM2 vs. PM1)	< 0.05	< 0.05	< 0.02
P value (PU vs. PR)	< 0.01	< 0.02	< 0.01

Discussion

A novel technique for increasing the size of the training dataset for various deep learning-based image analysis tasks requiring a large dataset to improve the model generalization and robustness was proposed. Our model is able to predict realistic multimodality (PET and CT) images based on a limited real dataset.

In contrast to previous studies, which used GANs for producing synthetic images, we aimed to use a small dataset containing *N* real patients and an analytical method called LB to produce a large synthetic dataset (2*N*+(*N*-1)) containing real and synthetic dataset. In our study, we had 35 real patients and created a total of 1225 images (35 real + 1190 synthetic). Contrary to machine learning approaches for generating data, the proposed technique does not need training, optimization, and training dataset.

Furthermore, another limitation of these approaches based on generative models was addressed by the proposed technique. GANs face challenges to learn all the underlying patterns/structures existing in certain types of imaging modalities or anatomical regions when the training dataset is limited and does not cover sufficiently large anatomical/functional variabilities. In this regard, the predicted images would only reflect or repeat the existing patterns/structures and anatomical/functional variabilities already exist in the original/training dataset. Another downside of GAN models is that the synthesized images do not have a reference for calculating the bias and discriminate between realistic synthesized images from noisy or skewed ones. The extended datasets available in the synthesized images through the proposed method helped the convolutional network to better decode the underlying features, thus resulting in superior performance. The convolutional network trained based on GAN generator applied simplistic noise reduction, thus leading to blurred, highly smoothed and slightly biased FD images

Robust-Deep could potentially be employed for any dataset wherein a template can be created/used to map the existing dataset to a common spatial coordinate. In this regard, the merging/sampling masks should be designed according to the imaging modalities, anatomical regions, and the endpoint of the study. Different mask designs can lead to different model performance, and it definitely helps to make the network more sensitive to the brain region of interest. For instance, when we plan to train a model for Alzheimer or Parkinson disease classification task, the mask can be more focused on the entorhinal cortex and hippocampus or caudate and putamen, respectively. Likewise, for a study focusing on the brain's gyrus, we can design a mask that selects different samples from the gyrus of different patients.

We involved both normal and abnormal patients to offer a heterogeneous representative dataset. Neurologic abnormalities present in our dataset included patients presenting with cognitive symptoms of possible neurodegenerative disease. Since the dataset for the training contained both normal and abnormal patients, it aided our method to produce various synthesized images which helped the network to avoid overfitting and guarantee robust and effective training. In this work, since all patients' images were registered to a single template, using data augmentation was not necessary and could not considerably change the results.

The generated diagnostic quality of ^{18}F-FDG brain PET images predicted from LD and LD+MR images corresponding to only 5% of the FD scan to evaluate the performance

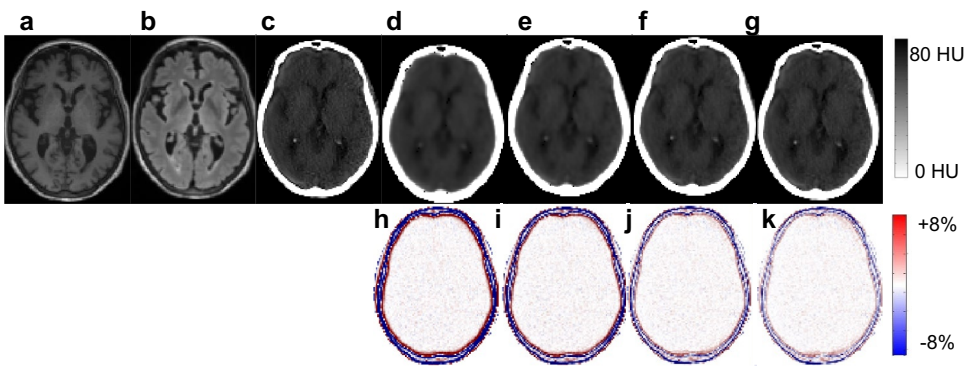

Fig. 8 The upper panel shows the reference and predicted images as well as bias maps. **a** T1-weighted MRI, **b** T2-weighted MRI, **c** reference CT, and predicted synthetic CT images generated using **d** PU, **e** PR, **f** PM1, and **g** PM2. The bias maps between original CT and predicted synthetic CT images generated using PU, PR, PM1, and PM2 are illustrated in **h**–**k**, respectively

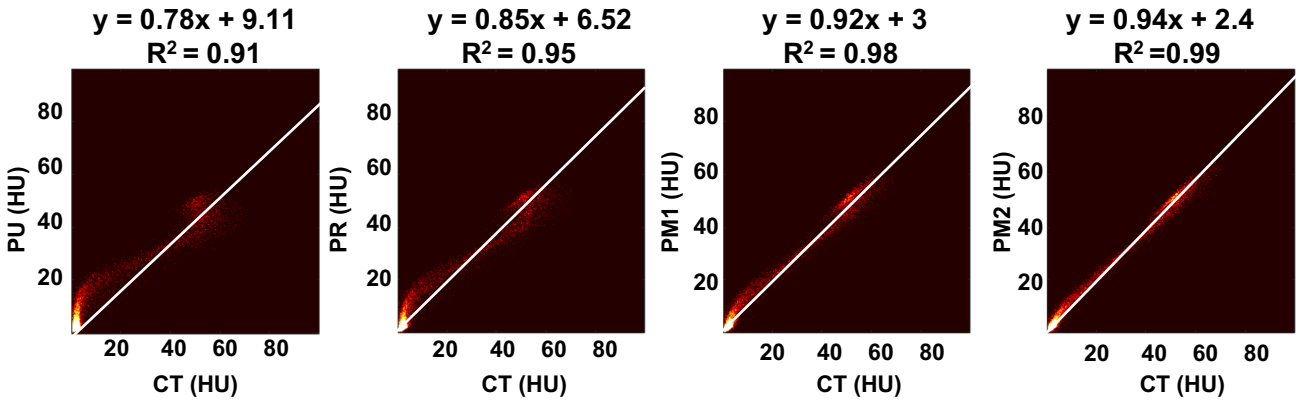

Fig. 9 Joint histogram analysis between reference CT and from left to right predicted synthetic CT images generated using PU, PR, PM1, and PM2 techniques, respectively

of both approaches (LD and LD+MRI) for estimation of FD PET images. It was shown that the synthesized FD images predicted from LD+MRI had a superior image quality and lower bias and variance compared to FD images predicted from only LD images. This highlights the value of employing anatomical MR images besides the LD images.

The results shown in Tables 2 and 3 indicate that if we train the DNN with only LD images plus synthesized images generated using the proposed DA method (PM2) improved the RMSE, SSIM, and PSNR metrics (0.20 ± 0.05, 0.93 ± 0.03, 37.14 ± 0.10, respectively) compared to the model trained using LD+MR images without using the synthesized dataset (PR) (0.24 ± 0.06, 0.90 ± 0.03, 32.96 ± 2.31, respectively), thus reflecting the effectiveness of the proposed DA technique when MR images are not available.

The direct generation of attenuation and scatter corrected PET images using deep learning approaches was previously investigated in the context of brain imaging [5, 7, 31]. In a study performed by Shiri et al. [5] for direct attenuation/scatter corrected brain images, 90 patients were used for training and reported SSIM and PSNR of 0.98 ± 0.006 and 39.2 ± 3.65, respectively. Robust and generalizable deep learning algorithms require large datasets in the training phase, which is impractical in real life or clinical setting. In the current study, we achieved SSIM 0.95 ± 0.06 and PSNR 31.08 ± 2.81 through employing image registration and data augmentation using only 35 patients, which resulted in comparable outcomes to previous studies. It is worth emphasizing that we cannot compare our results with previous studies because the images, data preparation, and DNN architecture are completely different. However, the proposed DA technique would potentially enhance the overall performance of similar deep learning models.

A major limitation of the proposed data augmentation technique is the requirement of image registration to a common spatial coordinate. In this regard, this technique is only applicable to imaging data corresponding to a specific anatomical region or organ. Nevertheless, a large number of applications in the field of medical image analysis focus on a specific organ or anatomical region, such as MRI-guided generation of synthetic CT for the pelvis [32, 33], head [9], and whole-body [34], standard/high dose CT or PET prediction [35], deep learning-based dosimetry [36, 37], PET attenuation and scatter correction in image space [7], metal artifact reduction [38, 39], various image-to-image transformation techniques [40], and dynamic imaging [41]. Though there might not be a standard template for some regions of the body or organs, the proposed data augmentation concept would work using images mapped to any reasonable common spatial coordinate. Another limitation is that the proposed data augmentation technique could not be employed for specific classification and survival rate estimation tasks in a straightforward manner. For such applications, the input and the annotated images (labels) should be processed/interpolated differently to create meaningful/realistic synthetic data.

Table 5 Comparison of results obtained from the analysis of image quality in predicted synthesized CT images (PU, PR, PM1, PM2) from MR images for the test dataset. SSIM: structural similarity index metrics, PSNR: peak signal to noise ratio, RMSE: root mean squared error

MRI to CT	RMSE	SSIM	PSNR
Unregistered predictions (PU)	0.34 ± 0.09	0.75 ± 0.06	17.78 ± 4.45
Registered predictions (PR)	0.29 ± 0.07	0.85 ± 0.07	21.78 ± 2.56
Registered mask#1 (PM1)	0.23 ± 0.07	0.92 ± 0.05	28.13 ± 2.01
Registered mask#2 (PM2)	0.22 ± 0.08	0.95 ± 0.02	30.55 ± 1.98
P value (PM2 vs. PM1)	< 0.05	< 0.02	< 0.05
P value (PU vs. PR)	< 0.02	< 0.05	< 0.05

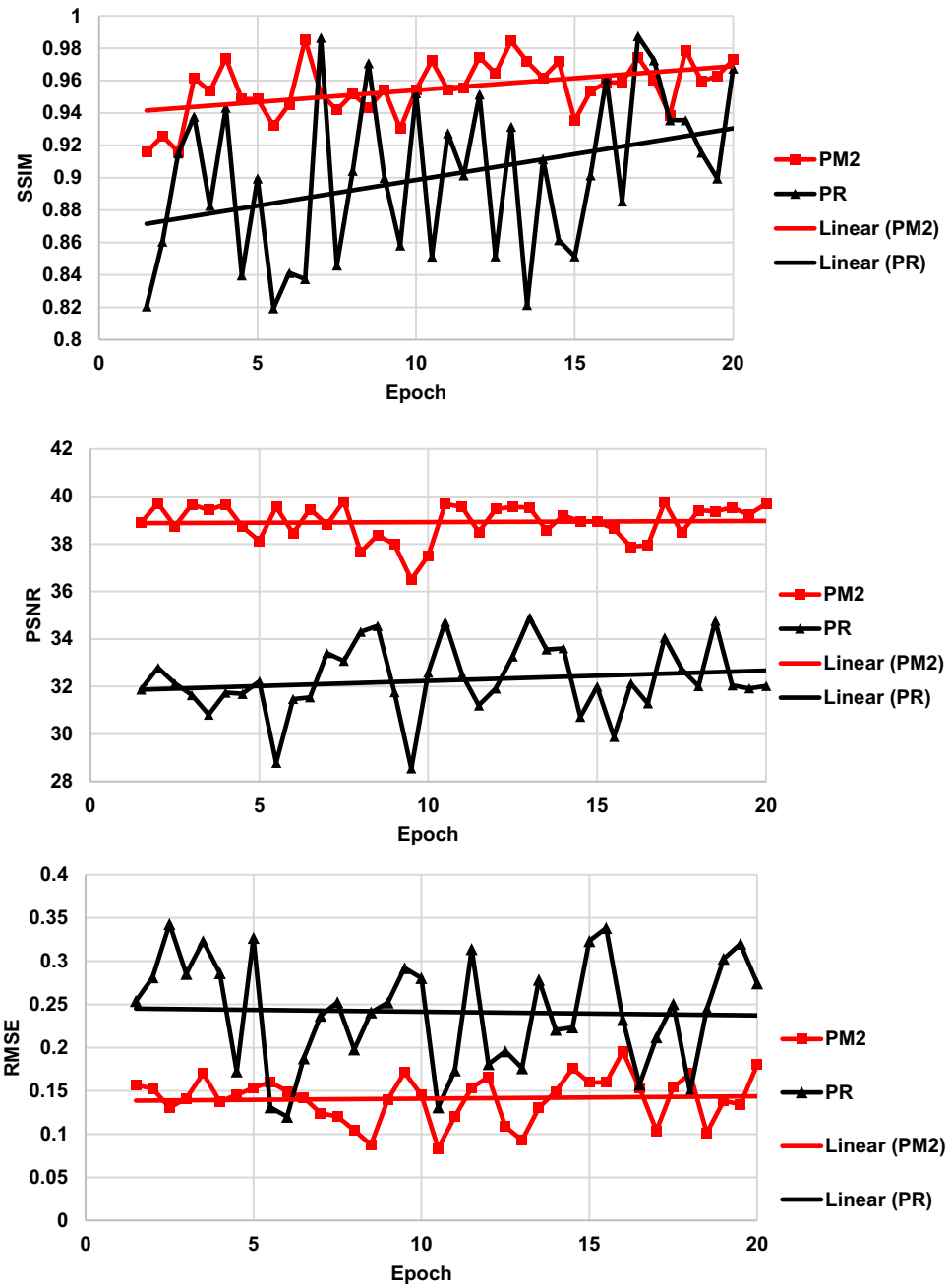

Fig. 10 Plots of SSIM, PSNR, and RMSE metrics versus epochs (1.5 to 20) calculated for LD PET + MRI to the FD PET translation task. The fluctuations and uncertainties of the metrics decreased noticeably after applying the proposed data augmentation technique

Conclusion

We have demonstrated that the DNN models' performance improved noticeably in four popular image-to-image translation tasks, including LD-PET to FD-PET, LD-PET + MRI to FD-PET, NAC-PET to MAC-PET, and MR to CT image conversions when exploiting the synthesized data created by our novel DA technique. The proposed DA model merges images of two different patients based on LB processing to generate a new realistic image. The qualitative and quantitative analysis proved that our model leads to superior performance, resulting in higher image quality and lower bias and variance compared to model training without using DA.

Acknowledgements This work was supported by the Swiss National Science Foundation under grants SNRF 320030_176052 and the Private Foundation of Geneva University Hospitals under Grant RC-06-01.

Availability of Data The dataset used in this work belong to the authors' institution and are not available to other parties.

Code Availability The code used in this work is available from the corresponding author upon request.

Data and Code Availability The Robust-Deep MatLab code is available from the authors upon request.

Declarations

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

Research Involving Human Participants All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Sanaat A, Arabi H, Mainta I, Garibotto V, Zaidi H: 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, 2020
2. Arabi H, Zaidi H: Deep learning-guided estimation of attenuation correction factors from time-of-flight PET emission data. Med Image Anal 64:101718, 2020
3. Shiri I, et al.: Ultra-low-dose chest CT imaging of COVID-19 patients using a deep residual neural network. Eur Radiol 31:1420-1431, 2021
4. Sanaat A, Shiri I, Arabi H, Mainta I, Nkoulou R, Zaidi H: Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging. Eur J Nucl Med Mol Imaging 48:2405-2415, 2021
5. Shiri I, et al.: Direct attenuation correction of brain PET images using only emission data via a deep convolutional encoder-decoder (Deep-DAC). Eur Radiol 29:6867-6879, 2019
6. Shiri I, et al.: Deep-JASC: joint attenuation and scatter correction in whole-body (18)F-FDG PET using a deep residual network. Eur J Nucl Med Mol Imaging 47:2533-2548, 2020
7. Arabi H, Bortolin K, Ginovart N, Garibotto V, Zaidi H: Deep learning-guided joint attenuation and scatter correction in multitracer neuroimaging studies. Human brain mapping 41:3667-3679, 2020
8. Arabi H, Koutsouvelis N, Rouzaud M, Miralbell R, Zaidi H: Atlas-guided generation of pseudo-CT images for MRI-only and hybrid PET–MRI-guided radiotherapy treatment planning. Phys Med Biol 61:6531-6552, 2016
9. Arabi H, Zeng G, Zheng G, Zaidi H: Novel adversarial semantic structure deep learning for MRI-guided attenuation correction in brain PET/MRI. Eur J Nucl Med Mol Imaging 46:2746-2759, 2019
10. Wang F, Casalino LP, Khullar D: Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med 179:293-294, 2019
11. Norgeot B, Glicksberg BS, Butte AJ: A call for deep-learning healthcare. Nat Med 25:14-15, 2019
12. Esteva A, et al.: A guide to deep learning in healthcare. Nat Med 25:24-29, 2019
13. Zhang L, et al.: When unseen domain generalization is unnecessary? rethinking data augmentation. arXiv preprint arXiv:190603347, 2019
14. Yu S, et al.: Robustness study of noisy annotation in deep learning based medical image segmentation. Phys Med Biol 65:175007, 2020
15. Rauber J, Brendel W, Bethge M: Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:170704131, 2017
16. Halevy A, Norvig P, Pereira F: The unreasonable effectiveness of data. IEEE Intell Syst 24:8-12, 2009
17. Noguchi S, Nishio M, Yakami M, Nakagomi K, Togashi K: Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. Comput Biol Med:103767, 2020
18. Liang D, Yang F, Zhang T, Yang P: Understanding mixup training methods. IEEE Access 6:58774-58783, 2018
19. Terrance V, Graham WT: Dataset augmentation in feature space. Proc. International Conference on Learning Representations (ICLR 2017): City
20. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321:321-331, 2018
21. Han C, et al.: Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. IEEE Access 7:156966-156977, 2019
22. HaoQi G, Ogawara K: CGAN-based Synthetic Medical Image Augmentation between Retinal Fundus Images and Vessel Segmented Images. Proc. 2020 5th International Conference on Control and Robotics Engineering (ICCRE): City
23. Bhattacharya D, Banerjee S, Bhattacharya S, Shankar BU, Mitra S: GAN-Based Novel Approach for Data Augmentation with Improved Disease Classification: Springer, 2020
24. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR: Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access 8:91916-91923, 2020
25. Schaefferkoetter J, Nai YH, Reilhac A, Townsend DW, Eriksson L, Conti M: Low dose positron emission tomography emulation from decimated high statistics: A clinical validation study. Med Phys 46:2638-2645, 2019
26. Della Rosa PA, et al.: A standardized [18F]-FDG-PET template for spatial normalization in statistical parametric mapping of dementia. Neuroinformatics 12:575-593, 2014
27. Shen J, Zhao Y, Yan S, Li X: Exposure Fusion Using Boosting Laplacian Pyramid. IEEE Trans Cybern 44:1579-1590, 2014
28. Ho Lee J, Choi I, Kim MH: Laplacian patch-based image synthesis. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): City
29. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T: On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. Proc. International Conference on Information Processing in Medical Imaging: City
30. Schoonjans F, Zalata A, Depuydt C, Comhaire F: MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 48:257-262, 1995
31. Yang J, Park D, Gullberg GT, Seo Y: Joint correction of attenuation and scatter in image space using deep convolutional neural networks for dedicated brain 18F-FDG PET. Phys Med Biol 64:075019, 2019
32. Arabi H, et al.: Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region. Med Phys 45:5218-5233, 2018
33. Bahrami A, Karimian A, Fatemizadeh E, Arabi H, Zaidi H: A new deep convolutional neural network design with efficient learning capability: Application to CT image synthesis from MRI. Med Phys 47:5158-5171, 2020
34. Dong X, et al.: Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging. Phys Med Biol 64:215016, 2019
35. Zhou L, Schaefferkoetter JD, Tham IWK, Huang G, Yan J: Supervised learning with cyclegan for low-dose FDG PET image denoising. Med Image Anal 65:101770, 2020

36. Lee MS, Hwang D, Kim JH, Lee JS: Deep-dose: a voxel dose estimation method using deep convolutional neural network for personalized internal dosimetry. Scientific reports 9:10308, 2019
37. Akhavanallaf A, Shiri I, Arabi H, Zaidi H: Whole-body voxel-based internal dosimetry using deep learning. Eur J Nucl Med Mol Imaging 48:670-682, 2021
38. Arabi H, Zaidi H: Truncation compensation and metallic dental implant artefact reduction in PET/MRI attenuation correction using deep learning-based object completion. Phys Med Biol 65:195002, 2020
39. Zhang Y, Yu H: Convolutional Neural Network Based Metal Artifact Reduction in X-Ray Computed Tomography. IEEE Trans Med Imaging 37:1370-1381, 2018
40. Jin CB, et al.: Deep CT to MR Synthesis Using Paired and Unpaired Data. Sensors (Basel) 19:2361, 2019
41. Wang C, et al.: Toward predicting the evolution of lung tumors during radiotherapy observed on a longitudinal MR imaging study via a deep learning algorithm. Med Phys 46:4699-4707, 2019
42. Sanaat A, Mirsadeghi E, Razeghi B, Ginovart N, Zaidi H: Fast dynamic brain PET imaging using stochastic variational prediction for recurrent frame generation. Med Phys 48(9) 5059-5071, 2021
43. Sanaat A, Shooli H, Ferdowsi S, Shiri I, Arabi H, Zaidi H: Deep-TOFSino: A deep learning model for synthesizing full-dose time-of-flight bin sinograms from their corresponding low-dose sinograms. NeuroImage 245:118697, 2021