IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 22, 2021, accepted October 19, 2021, date of publication October 22, 2021, date of current version November 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122194

Learning to Denoise Gated Cardiac PET Images Using Convolutional Neural Networks

JOAQUIN RIVES GAMBIN[1], MOJTABA JAFARI TADI[1,2], JARMO TEUHO[3,4], RIKU KLÉN[3,4], JUHANI KNUUTI[3,4], JUHO KOSKINEN[1], ANTTI SARASTE[3,4,5], AND EERO LEHTONEN[1], (Member, IEEE)

[1]Department of Computing, University of Turku, 20500 Turku, Finland
[2]Faculty of Engineering and Business, Turku University of Applied Sciences, 20520 Turku, Finland
[3]Turku PET Centre, University of Turku, 20520 Turku, Finland
[4]Turku PET Centre, Turku University Hospital, 20520 Turku, Finland
[5]Heart Centre, Turku University Hospital, 20521 Turku, Finland

Corresponding author: Joaquin Rives Gambin (joaquin.j.rives@utu.fi)

This work was supported by the Academy of Finland Through the MinMotion Project under Grant 314483.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the Hospital District of the South-Western Finland under Application No. ETMK 44/180/2012.

ABSTRACT Noise and motion artifacts in Positron emission tomography (PET) scans can interfere in diagnosis and result in inaccurate interpretations. PET gating techniques effectively reduce motion blurring, but at the cost of increasing noise, as only a subset of the data is used to reconstruct the image. Deep convolutional neural networks (DCNNs) could complement gating techniques by correcting such noise. However, there is little research on the specific application of DCNNs to gated datasets, which present additional challenges that are not considered in these studies yet, such as the varying level of noise depending on the gate, and performance pitfalls due to changes in the noise properties between non-gated and gated scans. To extend the current status of artificial intelligence (AI) in gated-PET imaging, we present a post-reconstruction denoising approach based on U-Net architectures on cardiac dual-gated PET images obtained from 40 patients. To this end, we first evaluate the denoising performance of four different variants of the U-Net architecture (2D, semi-3D, 3D, Hybrid) on non-gated data to better understand the advantages of each type of model, and to shed more light on the factors to take in consideration when selecting a denoising architecture. Then, we tackle the denoising of gated-PET reconstructions, revising challenges and limitations, and propose two training approaches, which overcome the need for gated targets. Quantification results show that the proposed deep learning (DL) frameworks can successfully reduce noise levels while correctly preserving the original motionless resolution of the gates.

INDEX TERMS Artificial intelligence, convolutional neural networks, biomedical imaging, cardiac gating, deep learning, positron emission tomography, respiratory gating.

I. INTRODUCTION

Positron emission tomography (PET) is the state-of-the-art imaging modality for various oncologic and cardiac imaging applications [1]. However, PET imaging is very susceptible to noise and motion artifacts, typically due to conscious, respiratory, and cardiac movements [2]. These lead to blurring and distortion in the images that can interfere in diagnosis and result in inaccurate interpretations.

The associate editor coordinating the review of this manuscript and approving it for publication was Hiram Ponce.

Motion correction techniques based on gating are proven to be effective in reducing motion blurring artifacts [3]–[5]. Gating in PET imaging consists in dividing the collected data into different groups or subsets (''gates'') by correlating the acquisition time with its corresponding amplitude or phase of the respiratory or cardiac motion [2], [4], [6]. For example, in device-driven dual gating a Real-time Position Management (RPM) system (Varian Medical Systems, Inc., Palo Alto, CA, USA) can be used to track longitudinal respiratory motion, while electrocardiography (ECG) is used to measure the cardiac electrical activity [5]. Dual gating considers the

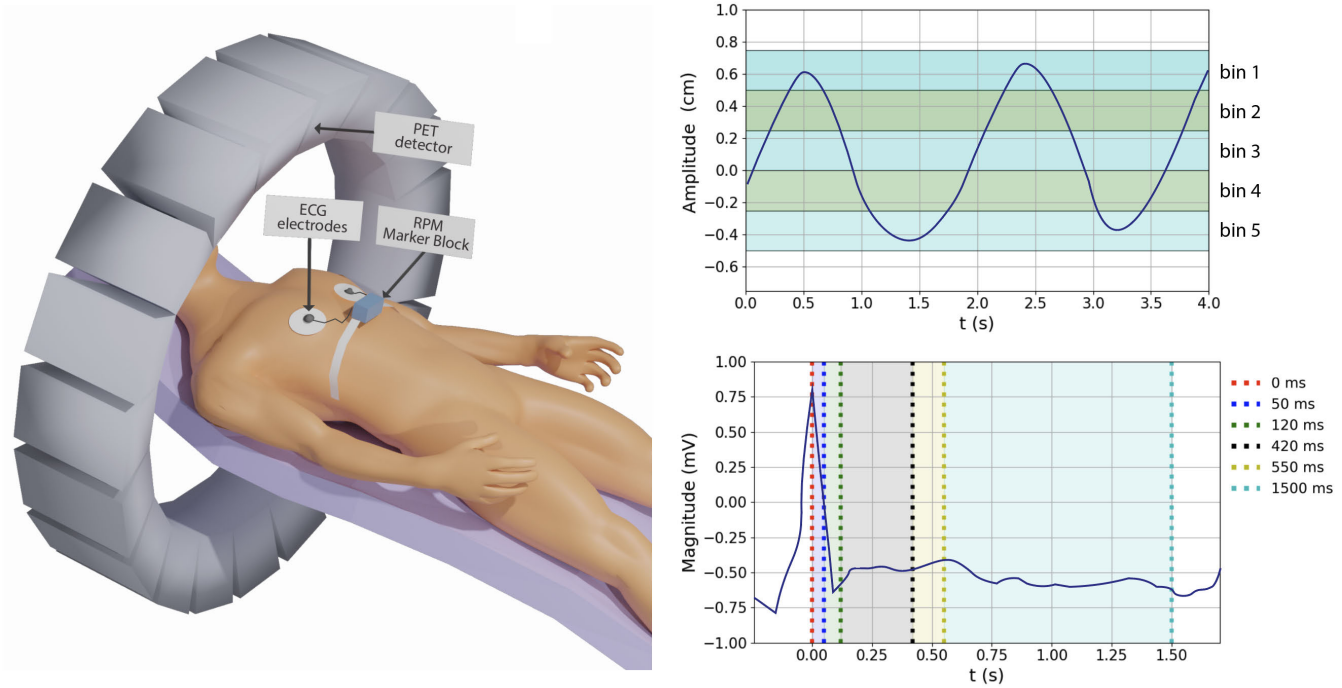


FIGURE 1. Left: General schematic of the set up used for gated-PET data acquisition, including an RPM marker block and ECG measurement electrodes. Right: Schemes of the respiratory (top-right) and ECG (bottom-right) dual-gating protocol. Respiratory bins (n = 5) were defined based on amplitude per breathing cycle, whereas cardiac bins (b = 5) were defined based on time from R-peak using non-equidistant fixed-time intervals.

different phases of cardiac or respiratory motion, and divides the list-mode PET data into specific bins based on their overlapping phase or amplitude in time. Each of the subsets can then be used to reconstruct an image dataset, which should have greatly reduced motion blur as compared to the non-gated image. Figure 1 illustrates the measurement setup used in the dual-gating PET data acquisition process. The RPM system includes an infrared (IR) camera which captures reflected beams from a marker block placed on the chest or abdomen of the subject.

Despite enhancing spatial resolution, the division of the data into multiple bins comes with a subsequent cost in image quality as only a fraction of the data is available for image reconstruction, resulting in increased levels of noise and thus image degradation. Reconstruction of images of reduced quality can likewise lead to quantitative inaccuracy and complicate clinical interpretation [4]. Thus, elevated noise levels of gated PET images represent a serious problem, currently holding back these gating techniques from being widely used in practice.

Existing conventional post-reconstruction denoising techniques include: Gaussian filtering, Non-Local Mean filtering (NLM) [7], [8], anisotropic diffusion [9], [10] and block-matching 3D (BM3D) [11], [12]. Variational PDE (partial differential equation) [13], one of the most recent techniques, have also gained popularity in many applications, such as image denoising [14] and segmentation [15]. However, deep learning could be a more effective technique to tackle these problems. Neural network's capability to learn non-linear complex relationships from data comes highly convenient in medical imaging denoising, where the statistical characteristics of noise are complex and difficult to model mathematically. Compared to traditional methods, DL models have the ability to model higher level features, and to integrate inter-patient information. Deep learning techniques have been extensively studied with promising performance in many medical imaging applications, such as reconstruction [16], [17], segmentation [18]–[20] and denoising [16], [17]. Recently, denoising methods based on convolutional neural networks, such as deep auto-context CNN [21], U-Net [22], and Generative Adversarial Networks (GANs) [23], have also been applied to non-gated low dose PET image achieving superior performance compared to conventional methods [24].

However, denoising of gated-PET images presents additional challenges. These methods usually require a high-resolution target or ground truth for training and evaluation of the denoising performance, whereas in gated-PET, high-dose targets are rarely available, except for the non-gated version, which intrinsically suffers from motion blurring. Only the work of Bo Zhou *et al.* [25] in 2020 tackles the denoising and motion estimation of gated-PET images using a Siamese Adversarial Network (SAN). Nevertheless, their proposed solution still depends on the availability of high-quality gated images.

Unsupervised denoising methods have also been studied in the literature, such as the Deep Image Prior [26]–[28]. However, the numerous limitations of the Deep Image Prior

approach – e.g. no clear stopping criteria or method to guarantee the consistency of the outputs, requires a new model to be fitted for each individual prediction, stochasticity of the results – make it rarely applicable in real practice. Other popular frameworks that can work without high-quality targets are usually based on the Noise2Noise (N2N) [29] approach, which will be further discussed in Section III-C. N2N based methods consist in pairing noise independent images, which still share the same underlying distribution [30], [31], or are highly correlated in structure [32].

In this work, we study the post-reconstruction denoising of cardiac dual-gated PET images using deep neural networks inspired by the U-Net architecture [18]. First, four U-net variants (2D, 2.5D, 3D, Hybrid) are trained on non-gated data for the purposes of selecting the best denoising architecture. Next, we apply the chosen denoising network (Hybrid) to the gated data in order to investigate whether the denoising capabilities of a neural network trained on non-gated scans can be extended to gated reconstructions, thus overcoming the need of high-quality gated targets. Finally, an additional training strategy based on the N2N [29] approach, was also investigated for comparison.

We show that our proposed denoising approach based on using non-gated low-count PET reconstructions as training data can successfully reduce noise levels while correctly preserving motionless resolution and anatomical details of the gated-PET images. It should be noted that the presented methods do not require additional or longer acquisition times, which broadens their utility.

II. MATERIALS AND METHODS

A. PET DATA

Our study consists of cardiac PET data from 40 coronary artery disease patients. The study protocol for clinical PET imaging was approved by the Ethical Committee of the Hospital District of the South-Western Finland (ETMK 44/180/2012). The subjects were scanned with the standard protocol used for vulnerable coronary plaque imaging at Turku PET Center on a GE Discovery 690 (D690) PET/CT system. The D690 is a fully 3D PET system containing a 64-slice Lightspeed CT system [33]. Demographic information of the subjects is described in Table 1.

Patients underwent contrast-enhanced coronary CT angiography (CTA) using the standard prospective ECG gated low dose CTA protocol with 50-100 ml of contrast agent (3.5 ml/s) and simultaneous acquisition of 64 parallel slices. After completion of the CTA study, a 3D PET scan of the heart was acquired in list-mode with ECG and respiratory gating with an acquisition time of 24 minutes. CTAC and CINE CT were acquired by using a low-dose CT with a tube voltage of 120 keV.

The 3D PET scans over the heart region were acquired in list-mode with ECG and respiratory triggering with the acquisition time of 24 minutes. 18F-fluorodeoxyglucose (^{18}F-FDG) was used as the radiotracer. Subjects were advised to keep their arms raised above the head by a supporting foam cushion to avoid truncation artifacts.

TABLE 1. Demographic information of the subjects.

	Clinical variable	Range
Sex (M / F)	$36/4$	-
Age (Y)	64 ± 9	$44 - 84$
Weight (kg)	86 ± 15	$47 - 116$
Height (m)	1.75 ± 0.09	$1.53 - 2.00$
Dose (MBq)	309 ± 26	$277 - 400$

Values are presented as number and mean±standard deviation.

B. ACQUISITION AND PROCESSING OF GATED-PET

For dual gating, 5 amplitude-based respiratory bins and 5 ECG-gated bins were used. Respiratory gating was performed using the exported respiratory curve from the RPM system. Cardiac gating (ECG-gating) was performed using 5 gating bins divided according to fixed-time intervals from the R-peak. The RPM system consists of a marker block, an infrared reflective marker placed on a plastic box positioned on the patient's thorax, and an infrared camera, which uses the IR beam reflected by the RPM marker block to track the respiratory cycle. For respiratory gating in PET and CT, amplitude gating was applied. The bins were divided equally by amplitude from end-inspiration to end-expiration. The gating thresholds were determined by equidistant sampling. The maximum threshold was defined as the mean plus one standard deviation of the amplitude maxima, whereas the minimum was determined from mean of the amplitude minima of the respiratory cycles. Only the cycles considered as ''valid cycles'' by the RPM system were used.

For cardiac gating, 5 bins using the cardiac triggers from list-mode data were determined. Non-equidistant, fixed-time gate assignment between subsequent R-peaks were used to define bins from end-systole to end-diastole. The time division to cardiac cycles was defined as 50 ms, 120 ms, 420 ms, 550 ms and 1500 ms (or until the next R-peak, whichever comes first) from the R-peak. Figure 1 illustrates our gating scheme based on the cardiac and respiration signals.

Static PET images were reconstructed using full 24 minutes of acquisition time. All gated and non-gated PET images were reconstructed with three-dimensional ordered subsets expectation maximization (3D-OSEM) reconstruction, using 2 iterations and 24 subsets. The reconstruction matrix size and field of view (FOV) were $256 \times 256 \times 47$ and 350 mm, respectively. A Gaussian post-filter of 6 mm was applied on all images. All quantitative corrections including decay, attenuation, scatter and randoms were applied to the reconstructed images.

C. DENOISING PROBLEM

In PET, images are reconstructed from a set of independent detector events, or count data. The acquired data is assumed to follow a Poison's distribution due to the stochastic decay process of the radiotracer, which corrupts the

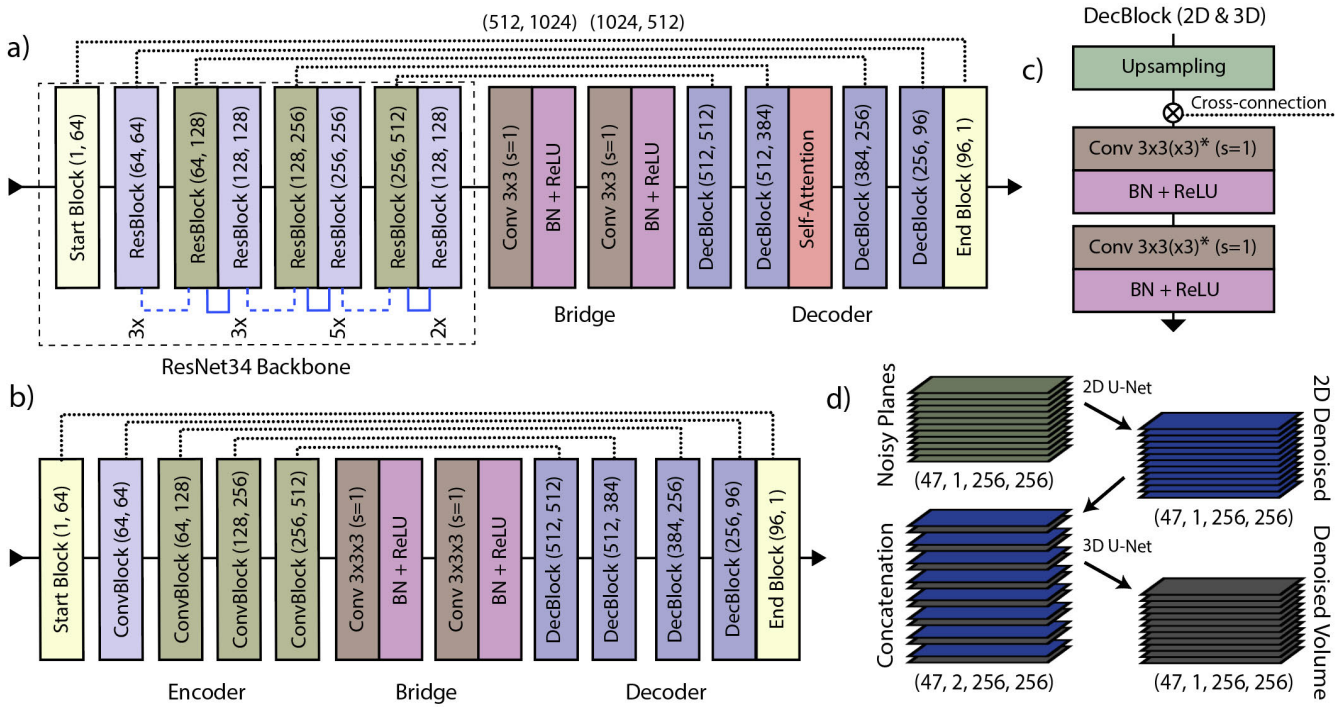


FIGURE 2. Illustration of the 2D (a), 3D U-Net (b) architectures, the decoder convolutional block (c) used in both architectures, and a scheme of the denoising process of a PET volume by the hybrid network (d). Dotted-lines in the architectures describe U-Net cross-connections between encoder and decoder. Residual connections between ResBlocks are represented by blue lines and dashed blue lines. On the contracting residual connections (blue dashed-line) an average pooling layer (stride = 2, kernel = 2) followed by a convolutional layer (1 × 1) is applied to the feature maps to match dimension of the main path. The input and output channels (m,n) of each layer is indicated between parenthesis. The semi-3D (or 2.5D) U-Net model follows the same base 2D (a) architecture but using three-channel input data. (×3)*: Only used in the 3D architecture (3 × 3 × 3 kernel dimensions).

reconstructed image. Thus, the corrupted input $\hat{x} \sim p(\hat{x}|y)$ is assumed to be a random variable distributed according to the clean target. In a low-count data set, the noise corruption is more severe due limited data counts.

Differently to traditional denoising methods, which usually require an explicit statistical modeling of the noise, NNs can learn the denoising function from existing data. Considering the corruption process (η) of an image x_i as

$$x_i = \eta(y_i), \tag{1}$$

our goal is to find a mapping function $f(\cdot)$ such as $f(x_i)$ is as close to y_i as possible. Using the MAE error, this is formulated mathematically as the following minimization problem:

$$\min_{H_f} ||H_f(x_i) - y_i||, \tag{2}$$

where in the mapping function $f(\cdot)$ is modeled by the neural network.

We can determine $f(\cdot)$ by training the network on a known dataset consisting of x_i (noisy) - y_i (clean) pairs such that Eq. (1) is optimized under the loss function minimizing the distance between input and target.

D. ARCHITECTURES

All the models were developed using the Pytorch (v1.5.0) [34] and Fastai (v2.0.0.18) [35] libraries. Training was carried out on a single NVIDIA V100 GPU with 32GB of memory.

1) 2D AND 2.5D U-NETS

In our 2D U-Net model as shown in Fig. 2.a, we use a modified version of the original residual networks (ResNet-34) concept [36] as the encoder backbone, deploying some of the improvements introduced by He *et al.* [37] in 2015. These modifications include the replacement of the 7 × 7 convolutional kernel in the start block by three 3 × 3 layers for improved computational efficiency, and the application of average pooling followed by a stride 1 convolution instead of a stride 2 convolution on the residual connections of the contracting residual block. In addition, we apply the ResNeXt configuration introduced by Xie *et al.* [38] for increased cardinality in the backbone residual blocks, consisting of independent convolutions, which are then aggregated by a 1 × 1 convolutional layer, instead of a single convolutional path. Xie *et al.* [38] showed that by increasing the number of convolutional paths, we can increase the model capacity to learn more complex transformations with a minimum impact in model performance, as the number of additional trained parameters is small.

In a similar manner, the structure of the up-sampling path is composed of blocks, which first apply an up-sampling operation followed by two 3 × 3 CNN layers and their activations. Pixel Shuffling ICNR [39] is used as the up-scaling operation in order to prevent the generation of "checkerboard" artifacts on the output images. Instead of a simple deconvolution

operation, Pixel Shuffling ICNR performs a nearest neighbour resizing of the image in combination a series of convolution operations, random pixel translocations and a special weight initialization that have proven to be effective in preventing these undesirable patterns caused by standard deconvolution operations [39].

The Rectified Linear Unit (ReLU) is used as activation function in all blocks except for the last 1×1 output kernel which uses a linear activation. We use reflection padding [40] and the weights are initialized using the Kaiming random initialization [41]. U-Net dense connections between encoder and decoder concatenate the activation maps of the convolutional layers with kernel of stride 2 (contracting residual blocks) and the activation maps of the corresponding upscaling block of the decoder. In order to better model relationships between distant spatial regions in the image, we also added a self-attention module [42] at the third block of the decoder before the end. The scheme of the architecture is shown in Fig. 2.a.

In our 2.5D model, instead of individual slices (256×256) from the PET volume, we use three neighbouring planes as input ($3 \times 256 \times 256$) by adding the two most adjacent planes as additional channels. This network, usually referred in literature as semi-3D or 2.5D, uses extra spatial information without applying full 3D convolutions, thus having a lower computational cost and number of parameters. On the other hand, the amount of spatial information is limited to only the two more adjacent planes.

2) 3D AND 2D/3D-HYBRID U-NETS

In PET imaging, where the data is represented in 3D volume formats, the use of the 3D CNNs allows the network to have access to all the spatial context information available in the volume, as opposed to 2D CNNs. On the other hand, 3D networks have higher computational costs and a larger number of parameters that increases exponentially as the model gets deeper, and as a consequence, the memory costs limit in high degree the size and depth of 3D DCNNs. A larger number of trainable parameters also require of larger amounts of training data for optimization, which is rarely abundant in the medical field. Hybrid architectures that exploit the advantages of both, 2D and 3D convolutions, in a single model have been proposed [19] with promising results. Following a similar approach, we developed a hybrid model by combining our 2D U-Net architecture with a 3D U-Net counterpart.

Our 3D U-Net deploys a similar configuration to the 2D U-Net described above, using the same block configuration of 2 convolutional layers ($3 \times 3 \times 3$ kernel) followed by Batch Normalization (BN) and ReLU activation. Due to the memory costs (GPU) of 3D convolutions, no residual connections between encoder blocks are used, and the encoder part is simplified to only 5 convolutional blocks both in the encoder and the decoder, corresponding filters of size 32, 64, 128, 256, and 512.

As Fig. 2.d shows the input volume is first denoised slice-by-slice through the 2D counterpart of the hybrid network. The outputs of the 2D network are subsequently concatenated with the original input volume and then fed as input to the 3D model to obtain the refined final prediction. The 3D counterpart of the hybrid in this architecture can take full advantage of the spatial information of the volume while the 2D part assists on the denoising and alleviates the amount of data required. Hence, the combination of both structures allows for an efficient extraction of intra-slice and inter-slice features.

E. PREPROCESSING AND DATA AUGMENTATION

We calculate the mean and standard deviation over the training set to perform a z-score normalization of the PET volumes. The next step differs depending on the U-Net architecture. For the 2D network, which requires 2D input images, the volumes are decomposed on its individual planes. In the case of the 2.5D U-Net models, the volumes are also decomposed on individual planes but with the addition of the two most adjacent planes as additional channels so that the input shape is $256 \times 256 \times 3$. The 3D U-Nets take as input the full PET volume of shape $256 \times 256 \times 47$ (x, y, z).

Due to the limited amount of data available, we use data augmentations to avoid overfitting and to improve generalization. The transformations we use for data augmentation in our pipeline include random rotations (z axis, 0-360°, $p = 0.25$), flips (x and y axes, $p = 0.33$), and in/out zooming (z axis, 0.85%-1.15%, $p = 0.15$). The probabilities (p) and intensity of the transformations are randomized and applied dynamically on every generated training batch and epoch, which allows for applying arbitrarily many combinations of different augmentations. The parameters of the transformation were empirically selected. We deliberately avoided the use of deformation transforms that could result in unrealistic anatomical scans. Some of the used transformations might still result in unrealistic anatomical configurations, e.g. flipping left-right would put the heart on the opposite side of the chest relative to what is normally seen in the PET images. However, due to the small size of the dataset, these transforms helped in reducing overfitting and improving overall performance, especially on the 3D models.

After training, in the case of the 2D and 2.5D U-Net models, the volumes are reconstructed from the individually denoised planes before evaluation. Finally, the denoised PET volumes are denormalized using the original mean and standard deviation in order to recover the absolute activity values.

F. EVALUATION

Performance metrics are calculated as average of an 8-fold subject cross-validation (CV). All PET reconstructions of the subjects included in the test set are always excluded from training. For each fold, we follow the same pipeline as described above from start to end. In order to assess the denoising performance of the different networks objectively, we need to establish a set of quantitative metrics. For comparison against the target, we choose two widely adopted image quality metrics, the peak signal to noise ratio (PSNR)

and the mean structural similarity index measure (SSIM). These are two well-known objective image quality metrics that have been extensively used in the literature to measure image degradation, quality and information loss [43], [44].

1) PEAK SIGNAL-TO-NOISE RATIO

The term peak signal-to-noise ratio (PSNR) is the ratio between the maximum possible power of a signal and the power of noise that distorts the quality of its representation [45, Ch. 4, pp. 127-135]. It is defined as follows:

$$\mathrm{PSNR} = 20\log_{10}\left(\frac{\mathrm{MAX}_f}{\sqrt{\mathrm{MSE}}}\right) \tag{3}$$

where MAX_f is the number of maximum possible intensity levels (minimum intensity level is supposed to be 0 in an image) and the mean-squared-error (MSE) is given by:

$$\mathrm{MSE} = \frac{1}{xyz}\sum_{i=0}^{x-1}\sum_{j=0}^{y-1}\sum_{z=0}^{z-1}\|f(i,j,k) - g(i,j,k)\|^2 \tag{4}$$

where $f(i,j,k)$ and $g(i,j,k)$ refer to the pixel values at location i,j,k in the reference and reconstructed image respectively. As PET images are usually reconstructed without a predefined maximum voxel value for the radioactive intensity, before calculation of the PSNR we rescale every volume to the range 0–1 and set MAX_f equal to 1.

2) STRUCTURAL SIMILARITY INDEX MEASURE

The main limitation of the PSNR is that it only performs a numerical comparison without taking in consideration how the difference in values is perceived by the human eye. A high PSNR score does not always correlate with the best perceptual and textural quality at visual inspection. The structural similarity index measure (SSIM) [46], on the other hand, tries to measure and correlate the characteristic of an image that have the most noticeable impact the human eye. The factors that the SSIM takes into consideration are the loss of correlation, luminance distortion and contrast distortion. The SSIM is defined as

$$\mathrm{SSIM}(f,g) = l(f,g)\cdot c(f,g)\cdot s(f,g),$$

where

$$\begin{cases} l(f,g) = \dfrac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \\ c(f,g) = \dfrac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \\ s(f,g) = \dfrac{\sigma_{fg} + C_3}{\sigma_f\sigma_g + C_3} \end{cases} \tag{5}$$

where f and g refer to an image patch of the reference and reconstructed image respectively, and μ and σ are the mean and standard deviation of that image patch. The SSIM index is calculated using a sliding window over the whole image and the average of the resulting values is used as the final mean SSIM score between the 2 images. Its values range from 0 to 1, where 0 meaning no correlation and 1 meaning that the images are identical. C_1, C_2 and C_3 are small constants used to avoid division by zero. The reformulated equivalent equation can be expressed as

$$\mathrm{SSIM}(f,g) = \frac{(2\mu_f\mu_g + C_1)(2\sigma_{fg} + C_2)}{(\mu_f^2 + \mu_g^2 + C_1)(\sigma_f^2 + \sigma_g^2 + C_2)} \tag{6}$$

In our implementation we set a non-overlapping window of size 17×17 pixels with zero-padding on image edges. As the SSIM metric was originally designed to compare RGB images, after denormalization of the PET volumes we rescale the voxel values to the range 0–255 before its calculation. We use the recommended [46] values of $C_1 = 1 \times 10^{-4}$, $C_2 = 3 \times 10^{-4}$, and $C_3 = C_2/2$.

3) VOIs

In a typical medical imaging scenario, a large portion of the scan does not contain relevant or useful information for clinicians when rendering diagnostics or treatment decisions. In order to evaluate the model's performance on the critical areas, we calculate the mean absolute error (MAE) over the two selected volumes of interest (VOIs) surrounding the heart and thoracic regions. The VOI's dimensions are: 82.2 mm $\times$ 112.4 mm $\times$ 22.9 mm (VOI-1); 82.2 mm $\times$ 95.9 mm $\times$ 9.8 mm (VOI-2).

4) PROFILE ANALYSIS

A profile analysis was performed to quantify the reduction in motion and ensure that no additional blurring was introduced on the denoised gates. The profiles were manually selected from a single-plane image patch overlapping the opposing walls of the left ventricle.

The peaks of the aggregated voxel values across the profile were used to calculate the full width at half maximum (FWHM). The FWHM difference relative to the non-gated image was used as measurement for motion minimisation and spatial resolution. The relative FWHM difference was calculated as follows:

$$\Delta\% = (\mathrm{FWHM_G} - \mathrm{FWHM_{NG}})/\mathrm{FWHM_{NG}} \times 100\% \tag{7}$$

where $\mathrm{FWHM_{NG}}$ refers to the FWHM measured from the non-gated image and $\mathrm{FWHM_G}$ to the FWHM calculated from a gated image. As baseline for the half maximum location, we use the local minimum between the peak of the interventricular septum and the peak of the lateral wall of the left ventricle.

As the motion of the wall of the interventricular septum is relatively slow, the corresponding gated and non-gated profiles are quite similar (see Figs. 6 and 8). Thus only FWHM measurements of the right-most peaks of these figures, corresponding to the lateral wall of the left ventricle, were calculated and included in the profile Tables 3 and 4.

III. EXPERIMENTS

A. NON-GATED PET STUDY

In order to better understand the potential benefits of each model in denoising, we validate them by using the non-gated

PET data for training and evaluation, as it leads to deploy the full-count reconstructions as the ground truth. We train the networks on non-gated low-count reconstructions (5%, 10% and 14% list-mode data) as input, while using the images reconstructed using 100% of the list-mode data as targets. The part of the list-mode data used in the low-count reconstructions is randomly selected by dividing the total list-mode data points in 20, 10 and 7 groups of equal size. That is, for every subject we obtain 1 target (Y) and 37 training instances (X), all of which paired to the same target.

All models were trained for 15 epochs, using the Adam optimizer [47] with a weight decay of 1×10^{-3} and the MAE loss. The MAE loss was empirically selected based on preliminary experiments in which we evaluated four different losses (MAE, MSE, MAE-SSIM, MSE-SSIM). The MAE loss consistently yielded the best performance scores, and highest output quality at visual inspection. For the learning rate and momentum, we used the Leslie Smith's one cycle scheduling policy [48], which consists of 2 training phases. In phase 1 of the training cycle, we start training at minimum learning rate ($lr_{\max}/25$) and linearly increase it to $lr_{\max} = 1 \times 10^{-3}$, while letting the momentum to decrease from $m_{\max} = 0.95$ to $m_{\min} = 0.85$ linearly. In phase 2, the learning rates follows a cosine annealing from $lr_{\max}$ to 0, where the momentum goes from $m_{\min}$ to $m_{\max}$ with the same annealing.

No stopping criteria was used as higher number of epochs didn't show any significant worsening or improvements in performance. We kept 15 epochs on all models for equal training conditions between models, and to ensure that the number of epochs was not a limiting factor on any of the U-Net architectures. In order to improve stability during training and achieve a better convergence, the training of the hybrid U-Net was executed following the same method but in 3 different steps:

1) Fitting of the 2D part of the network independently for 1 cycle of 5 epochs.
2) Fitting of the 2D/3D hybrid network with the parameters of the 2D part fixed for 1 cycle of 5 epochs.
3) Unfreezing of the 2D part and training of the whole network jointly a lower learning rate ($lr_{\max}/10$) for 1 cycle of 5 epochs.

The 2D and 3D sub-models that compose the hybrid model use the configuration and architecture shown in Fig. 2. The scores reported for these models are always calculated over the 3D volumes (or VOIs) after reconstruction as indicated in the Section II-E to ensure the comparability of the results between models.

Quantitative comparison of the scores (Fig. 3) showed that the hybrid model outperforms the other models achieving the best performance in all three evaluation metrics. The 3D model, despite having access to extra spatial information, showed lower performance on the MAE and SSIM scores (106.5 and 0.664) as compared to the 2D U-Net model (105.7 and 0.685). Only on the PSNR score the 3D network ranked differently, achieving a slightly better score than the

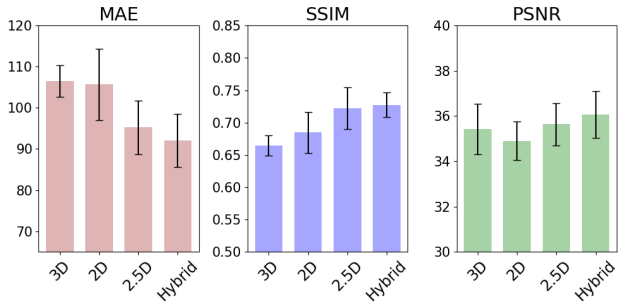


FIGURE 3. Performance comparison of the different U-Net models. The results are calculated from an 8-fold subject cross-validation.

TABLE 2. Average MAE scores of the selected VOIs around the heart and thoracic regions of two different test subjects. Location illustrated in Fig. 4.

	MAE	
	VOI-1	VOI-2
Gaussian	985.6	236.8
NLM	906.5	278.9
BM3D	997.5	335.4
Prior	981.3	274.4
3D	890.5	264.0
2D	626.7	186.4
2.5D	622.1	172.0
Hybrid	624.0	170.5

Units: Bq/ml.

2D network. Despite the 2D and 3D networks having a similar MAE score, the MSE of the 2D network (63.0 kBq/ml) is higher than the 3D network (54.2 kBq/ml), which is used in the calculation of the PSNR score. It suggests that the 2D network has a smaller overall error but with specific voxels containing large errors, which are amplified by the MSE. The MSE score of the 2.5D and hybrid models was consistent at 44.7 kBq/ml and 39.4 kBq/ml respectively.

We can also observe a clear overall improvement between the 2D and 2.5D U-Nets, with a reduction of the MAE error from 105.7 to 95.2, showing that adding neighboring slices as extra channels can significantly improve the denoising performance of the 2D-CNNs while avoiding the heavy memory/computation cost of the 3D-CNNs. In this particular case, the training of the 3D models took approximately 4 to 5 times longer than the 2D models.

The error calculated from the selected VOI's (Table 2) also showed a similar trend with 2.5D and hybrid models achieving the best MAE scores. All deep learning models outperformed the conventional methods in MAE score as well as in perceived visual quality as shown in Fig. 4.

Based on these results, the hybrid network was selected for the subsequent denoising experiments.

B. GATED-PET STUDY

In this section, we use the previous models to perform inference on gated PET data. Here, we must face a different

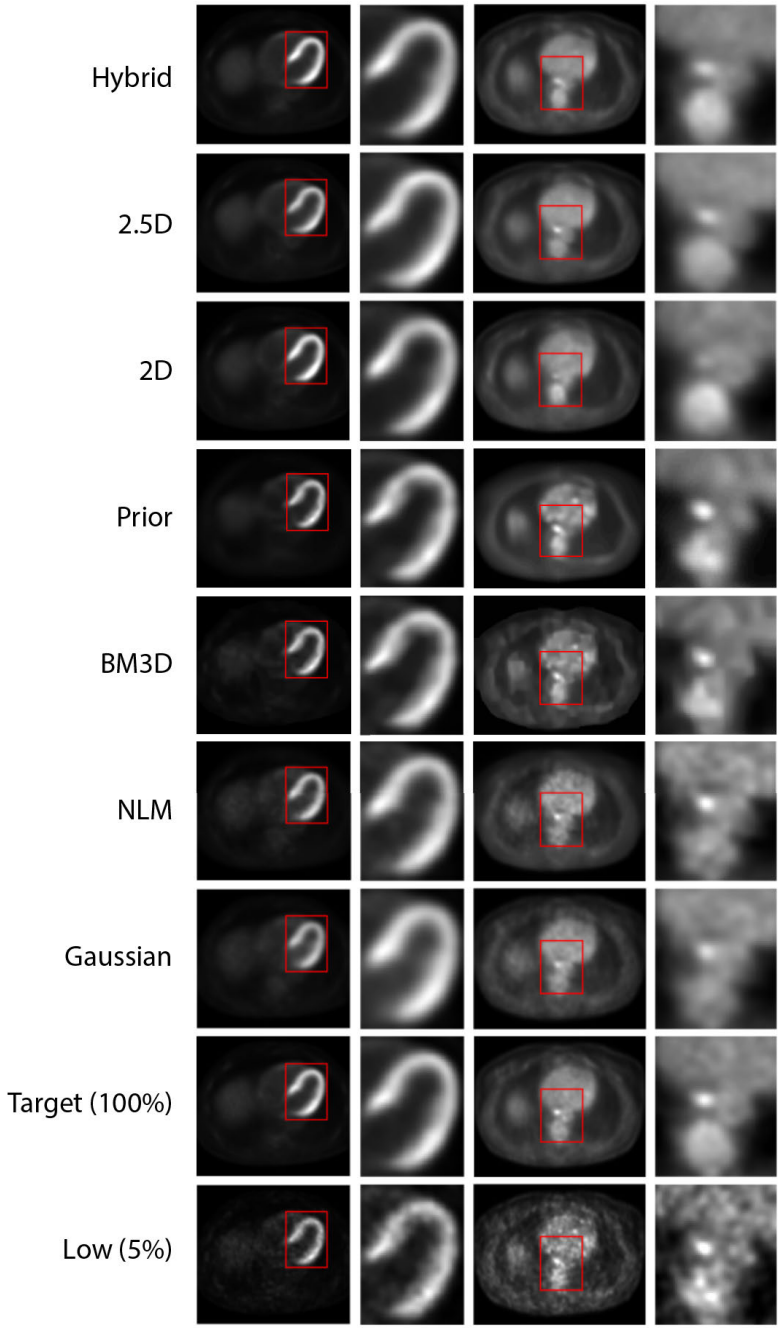


FIGURE 4. Non-gated low-count (5%), full-count (100%) and denoised images comparing the performance of three conventional methods (Gaussian and NLM filtering and BM3D), the Deep Image Prior and our 2D, 2.5D and Hybrid U-Nets. The zoom-in region indicated in red correspond to the VOIs 1 and 2 with MAE scores shown in Table 2. Gaussian filter σ: 2; NLM window: 5 × 5 × 5; NLM patch-distance: 13 pixels; BM3D σ: 20; Deep Image Prior: Author's original implementation [28], using the drunet-gray model with a noise level of 12.

scenario where we have 25 gates with a fraction of the total count-data that ranges between 1% and 15% depending on the gate. This means that the model should be able to account for a wider range of noise levels. This was indeed the first problem affecting model performance we encountered in our preliminary experiments. We noted that the denoising performance degraded when the considered gate contained a fraction of data that was significantly smaller than the fraction of the training set (only 10% reconstructions). However, this problem was partially solved after the addition of training instances of the same subjects reconstructed at different fractions (5%, 10% and 14%) of the total count data as described in Section III-A.

Another major challenge is the lack of a ground-truth to be used as target. In this case, the absence of a gated high-quality version to be used as a ground truth hampers the use of

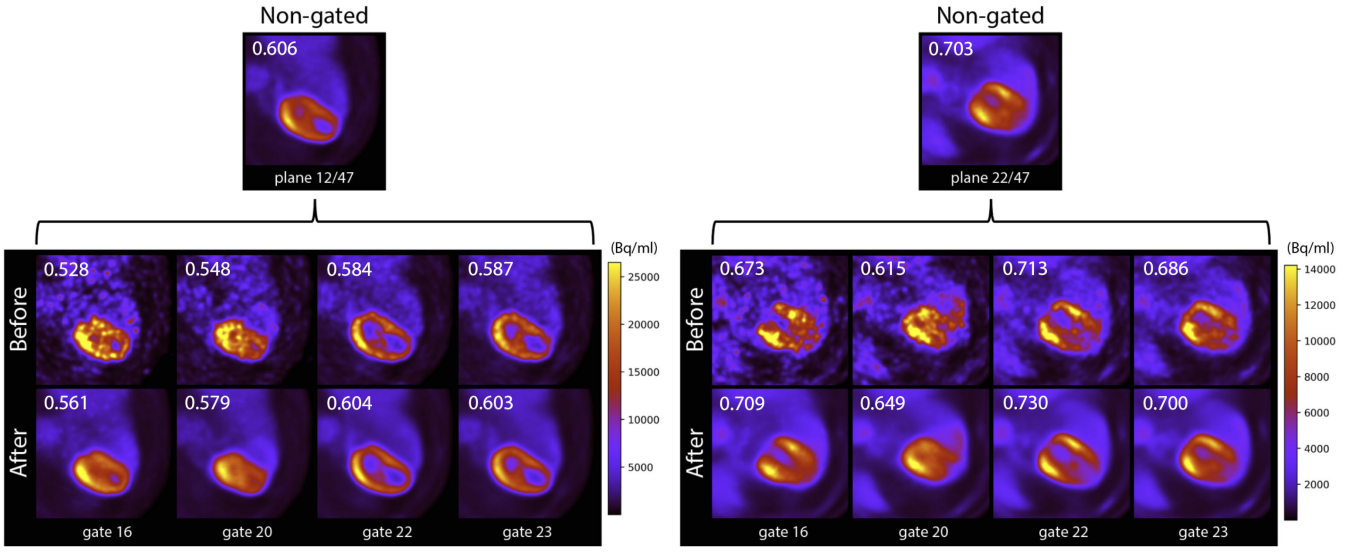


FIGURE 5. Examples of gated-PET reconstructions before and after denoising by the Hybrid U-Net model. The signal-to-noise ratio (SNR) is indicated on the top-left corner of each PET image. The indicated SNR was calculated as $I_{\text{mean}}/I_{\text{sd}}$, where I_{mean} and I_{sd} are the average and standard deviation uptake of the plane shown in the figure.

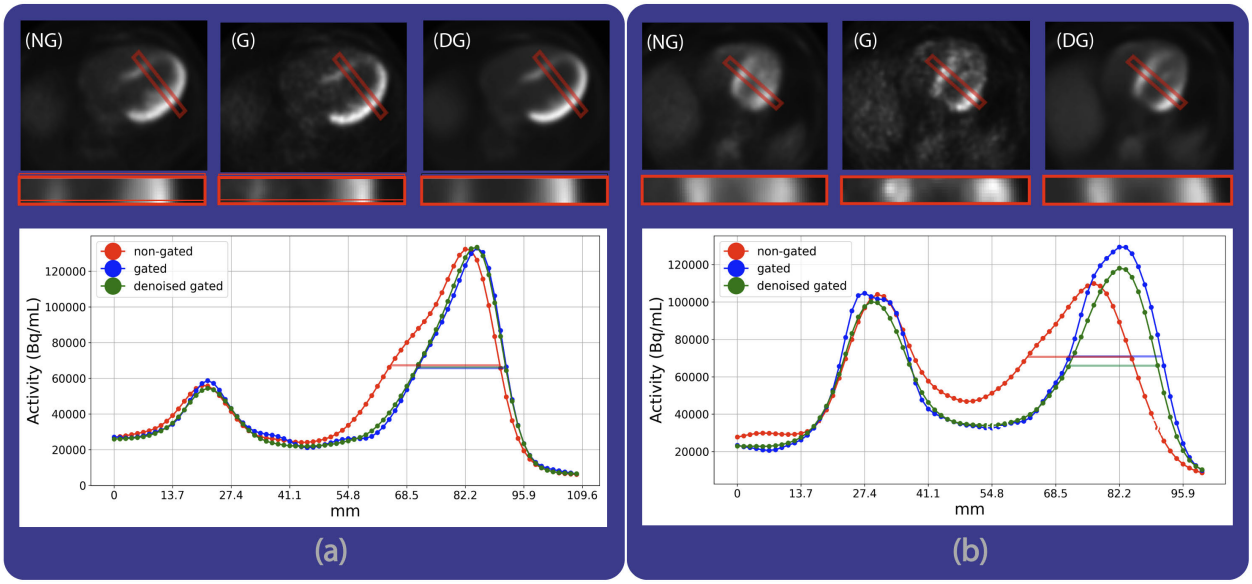


FIGURE 6. Profile analysis of the selected region (red box) in 2 different subjects. In the profile chart, the horizontal lines indicate the location of the half-maximum. FWHM measurements are shown in Table 3. Non-gated full-count image (NG); Gated image (G); Denoised gated image (DG).

the previous noise-to-clean (N2C) training approach. On the above premises, in this section we study the feasibility and performance of the networks trained using only non-gated low-count reconstructions as in the previous section when denoising also the gated reconstructions of the test subjects.

The results displayed in Fig. 5 and 6 demonstrate the feasibility of denoising gated PET images using models exclusively trained on non-gated low-count reconstructions. The learned denoising capability of the network was successfully transferred to the gated reconstructions of the test subjects (non-gated data was also completely excluded from training set) achieving to restore the quality of the images and effectively reducing noise without introducing observable side-effects or artifacts. As shown in Fig. 5, the noise of the gated-PET images due to the inherently lower data count has been significantly reduced while correctly preserving the motionless resolution, sharpness, and anatomical characteristics of the myocardial regions.

The profile analysis (Fig. 6 and Table 3) also indicate an improvement in image sharpness. The higher spatial resolution and lower motion blurriness of the gated images is still preserved after denoising. Furthermore, noise artifacts

TABLE 3. FWHM measurements of the profiles (a) and (b) shown in Fig. 6. The peaks selected for the measurements correspond to the right-most peak (lateral wall of the left ventricle).

	Profile A		Profile B	
	FWHM	Δ%	FWHM	Δ%
Non-gated	25.2	0%	24.8	0%
Gated	20.2	19.8%	19.5	21.3%
Denoised gated	19.5	22.6%	19.6	21.0%

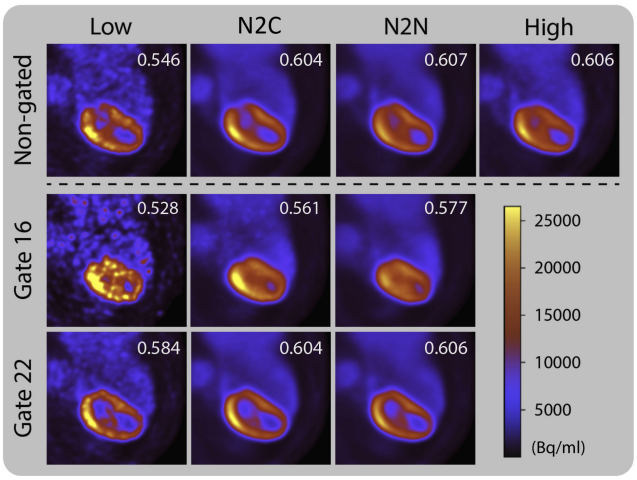


FIGURE 7. Examples of non-gated and gated outputs denoised by the hybrid model using the N2C and N2N training approaches. The signal-to-noise ratio (SNR) is indicated on the top-right corner of each PET image. The indicated SNR was calculated as $I_{\text{mean}}/I_{\text{sd}}$, where I_{mean} and I_{sd} are the average and standard deviation uptake of the plane shown in the figure.

present in the gated images, which yield fluctuations and small "bumps" in the profiles, have been corrected on the denoised outputs.

C. NOISE-TO-NOISE APPROACH TO PET DENOISING

Conventional image restoration tasks using DCNNs consist of pairing a noisy or low-quality image x_i to a clean target of higher quality or resolution y_i. However, recent studies suggest that same or even better results can be potentially achieved by simply pairing noisy images. This approach is called Noise-to-Noise (N2N) [29] as it does not require higher quality images to be used as targets, instead, noise realization of the same underlying distribution are considered as targets. In PET imaging the N2N approach is appealing, as it allows to use motion-mitigated gated images as targets instead of motion-blurred non-gated images.

As explained in II-C, the denoising of corrupted images using NNs is a minimization problem expressed as

$$\min_{H_f} ||H_f(x_i) - y_i||, \tag{8}$$

where x_i and y_i denote the noisy and clean reconstructed images, and the function $f(\cdot)$ represents the neural network. Unfortunately, the acquisition of clean targets to be used has ground truth is rarely possible in many clinical settings,

TABLE 4. FWHM measurements of the profiles (a) and (b) shown in Fig. 8. The peaks selected for the measurements correspond to the right-most peak (lateral wall of the left ventricle).

	Profile A		Profile B	
	FWHM	Δ%	FWHM	Δ%
Non-gated	22.3	0%	24.9	0%
Gated	16.2	27.3%	19.6	21.3%
Denoised gated	16.4	26.5%	19.7	20.9%
Denoised gated N2N	17.2	22.9%	19.6	21.3%

e.g. gated-PET. Moreover, It assumes that the corrupted input $\hat{x} \sim p(\hat{x}|y)$ is a random variable distributed according to y. However, it is not totally true, as even "clean" images, reconstructed from high count datasets, are inevitably affected by some level of noise. The hypothesis of the N2N approach states that it is possible to determine $f(\cdot)$ by using only noisy reconstructions. Following this approach, Eq. (8) optimization is reformulated as

$$\min_{H_f} ||H_f(x_i) - x_{ii}||, \tag{9}$$

where x_i and x_{ii} are two different noisy realizations of the same unobserved y_i.

In our dataset we have a high-count target y for each noise realization x at a lower percentage of total count data. In the classical approach where we pair all the noise realizations to the same target y, e.g. at 10%, a total of 10 pairs or training instances is obtained. In the N2N approach, however, each noise realization is paired against each other giving a total of 90 combinations. As long as the input and target are conditioned on the same underlying unobserved clean target, the loss minimization then becomes a maximum expectation problem that should model the real underlying distribution. Theoretically, the authors [29] showed that with infinite data Eq. (8) and Eq. (9) are equivalent, whereas under finite data, the variance of the estimate should be equal to the average variance of the corruptions in the targets, divided by the number of training samples.

In order to study the performance of the N2N approach on PET data, we train again the network as in Section III-A, but pairing within subject's low-count reconstructions with each other instead of using high-quality targets.

Results showed that the N2N approach can effectively remove noise and motion artifacts achieving comparable results to the N2C approach in non-gated as well as in gated reconstructions. The profile analysis of the N2N outputs (Fig. 8) are also consistent with the profiles obtained following the N2C approach. The relative FWHM measurements indicated that both training approaches achieved similar motion reduction levels.

Particularly, we observed the most substantial improvement in scans with higher levels of noise, such as the gates with lower percentage of data available for image reconstruction, or the first and last volume slices where the loss of sensitivity of the detector causes significant levels of noise even in the high-quality targets reconstructed using 100% of

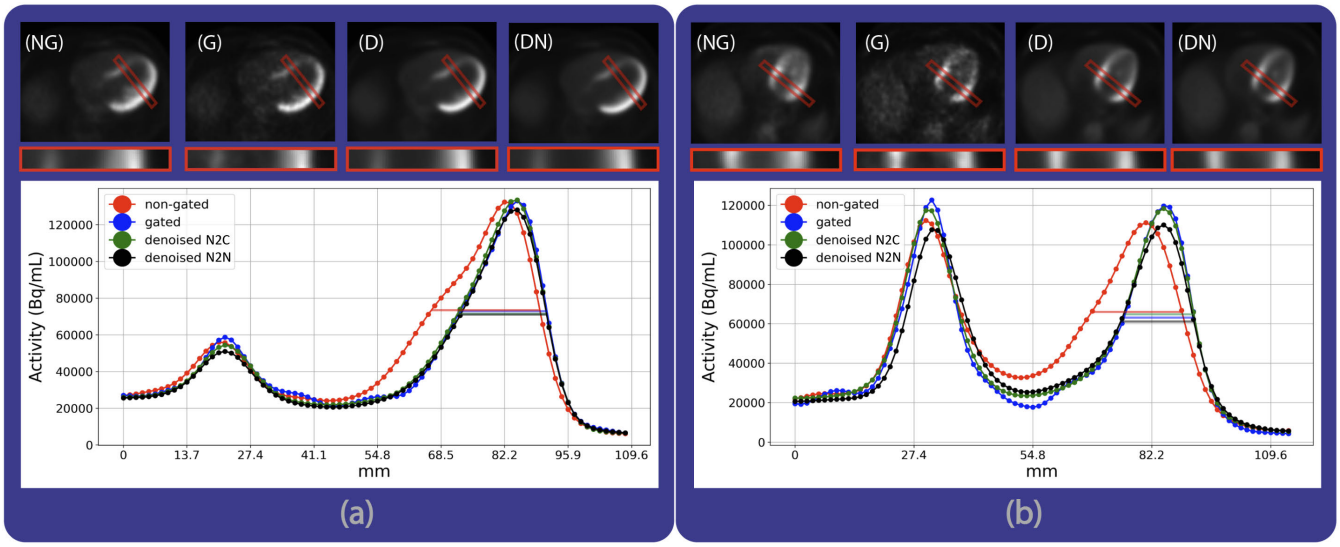


FIGURE 8. Profile analysis of the selected region (red box) in 2 different subjects. In the profile chart, the horizontal lines indicate the location of the half-maximum. FWHM measurements are shown in Table 4. Non-gated full-count image (NG); Gated image (G); Denoised gated image (D); N2N denoised gated image (DN).

the count data. When using the N2C approach, noise artifacts present in the targets are often reproduced by the model on the outputs. This supports the hypothesis that the N2N approach could potentially be a good solution in the absence of good targets.

IV. DISCUSSION

In the quantitative comparisons of Section III-A it was seen that the hybrid model yields the best overall performance followed closely by the 2.5D model, whereas the 3D model showed a significantly lower performance despite having access to further spatial information. Two factors related to the limitations of 3D convolutions could explain it. First, we have the computational cost and memory allocation of 3D convolutions, which in our case (32GB of GPU memory available) limited the size of the 3D models to only 28 convolutional layers as opposed to the 56 layers of the 2D models. The second factor is the higher number of the training parameters of 3D convolutions. Despite having less convolutional layers, the number of the parameters of the 3D model is approximately 4 times higher at 172M as opposed to the 41M parameters of the 2D model. In ideal situations and with enough training data, such high number of parameters could be advantageous and would enable a better modelling. However, it also supposes an optimization burden, demanding of higher amounts of training data, condition that clearly could not be fulfilled with the present training dataset (35 subjects). In the hybrid model, this burden is partly alleviated as the outputs of the 2D counterpart can effectively assist the 3D counterpart during the training phase facilitating a faster and more efficient convergence.

The above hypothesis is supported by the experimental results shown in Fig. 9 where the represented model performance versus amount of training data shows that with limited training data the 2.5D U-Net is favoured. Figure 9 also suggest that the improvement achieved by the hybrid network over the 2D networks could be more significant in conditions where the number of training instances available for training is higher. Although the 2D CNN achieved similar scores as compared to the hybrid network, extra training data did not yield any significant improvements. Due to the higher complexity and more parameters of the 3D U-Net, larger number of training datasets is indispensable to achieve a stable state of the trained network. The PSNR and SSIM curve on Figure 9 showed that the 2D and 2.5D U-Net already reached the state of convergence when 34 subjects were used for training while the curve of 3D U-Net was still in the rising state. This suggested that data from more patient might be required for the 3D U-Net training. Another insight to extract from this experiment is that the size of the training dataset should be taken into consideration when selecting the denoising network.

An additional advantage of network architectures that use extra adjacent slices for denoising, such as the 3D models, and even the 2.5D models, is the noticeable higher visual quality in coronal and sagittal perspectives. These perspectives specially benefit of the extra context information as it helps the network in achieving a better inter-slice consistency in the reconstructed output volumes. A clear example of this side-effect can be seen in Fig. 10.c, where the lack of inter-slice consistency of the 2D U-Net results in a pronounced blurriness of the coronal view.

In the gated-PET study, we hypothesized that the denoising capabilities of a network trained on non-gated data by mapping high quality targets to downsampled reconstructions can be extended to gated reconstructions, overcoming the need

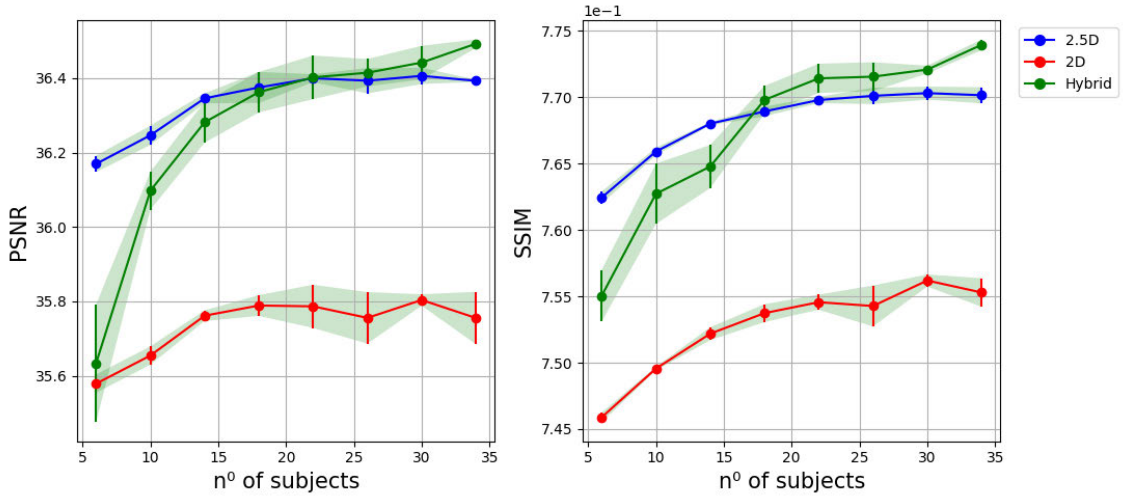


FIGURE 9. PSNR and SSIM scores on a 6 subjects test set versus the number of subjects available for training, each include 37 low-count reconstructions as described in Section III-A. Evaluation results are represented as the average (dots) and 95% confidence interval (vertical lines) of the scores obtained by each model after 5 repetitions of the same experiment.

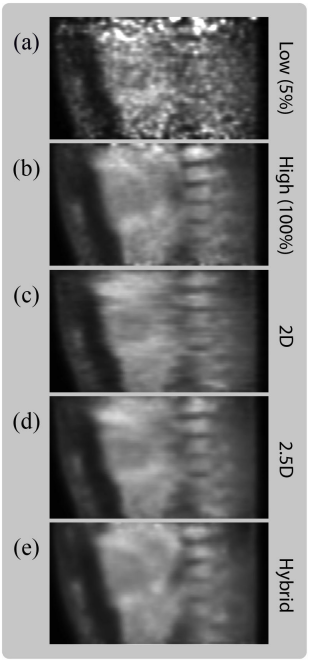


FIGURE 10. Coronal view of a subject illustrating the higher quality and consistency between slices achieved by the 2.5D (d) and hybrid (e) networks, as opposed to the 2D (c) network.

of high-quality gated targets. When following this approach, we assume the characteristics of the noise to be similar in gated and non-gated images. The results in Figs. 5 and 6 show a successful restoration of the images quality and effective reduction of noise levels without introducing observable side-effects or artifacts in the gated outputs. However, further research is needed to ensure that such assumption is safe and verify that any possible effects caused by distribution changes between non-gated images used for training and gated images at the time of inference will not cause major effects in performance. In future, we intend to re-assess in depth this assumption and measure any possible effects in performance. For example, high quality gated reconstructions could be used as evaluation targets against the denoised outputs. Such an approach requires longer PET acquisition times and radiotracer dosages for acquiring the high quality gated images, and could be implemented for example as a phantom study.

In Section III-C, results showed that the N2N approach can also be applied to gated-PET as an alternative solution to overcome the need of gated targets or even achieve higher denoising performance, especially when the performance of the models is limited by the low quality and/or presence of noise in the targets. Nevertheless, since this is not an unsupervised method, it still requires of additional data acquired under similar conditions and/or noise realizations for training. A method consisting in the use of bootstrap resampling [49] for the artificial generation of extra noise realization in PET has been proposed as a possible solution to alleviate this problem and would be worth further research [50]. Taking into consideration all factors and limitations of the N2N method, it might not offers much benefits for normal PET scans, however, it still makes sense for gated images where the noise levels are significantly higher than in normal non-gated scans.

V. CONCLUSION

We presented a deep learning post-reconstruction denoising study for cardiac gated-PET images using patient data, and considered solutions that do not require simulated ground truths or gated targets for training. The results showed that the proposed models can successfully reduce noise levels while correctly preserving the motionless resolution and anatomical characteristics of the gates. We applied and evaluated

different networks on both, gated and non-gated scans, providing further evidence of the denoising performance of U-Net based architectures. The results also showed the benefits of a hybrid solution in addressing the limited access to spatial context information of the 2D DCNNs and the high data requirements of 3D DCNNs. Finally, we showed the effectiveness of the N2N approach in restoring the underlying activity distribution of the gates, indicating that N2N could be a practical solution in the absence of high-quality targets.

ACKNOWLEDGMENT

The authors would like to thank the support and computational resources facilitated by the CSC-Puhti super-computer, a non-profit state enterprise owned by the Finnish state and higher education institutions in Finland. Mojtaba Jafaritadi is thankful of Ulla Tuominen Foundation and State Research Funding of Turku University Hospital.

REFERENCES

[1] J. M. Ollinger and J. A. Fessler, "Positron-emission tomography," *IEEE Signal Process. Mag.*, vol. 14, no. 1, pp. 43–55, Jan. 1997.

[2] S. A. Nehmeh, Y. E. Erdi, K. E. Rosenzweig, H. Schoder, S. M. Larson, O. D. Squire, and J. L. Humm, "Reduction of respiratory motion artifacts in PET imaging of lung cancer by respiratory correlated dynamic PET: Methodology and comparison with respiratory gated PET," *J. Nucl. Med.*, vol. 44, no. 10, pp. 1644–1648, 2003.

[3] A. Martinez-Möller, D. Zikic, R. M. Botnar, R. A. Bundschuh, W. Howe, S. I. Ziegler, N. Navab, M. Schwaiger, and S. G. Nekolla, "Dual cardiac–respiratory gated PET: Implementation and results from a feasibility study," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 34, no. 9, pp. 1447–1454, Aug. 2007.

[4] F. Büther, M. Dawood, L. Stegger, F. Wübbeling, M. Schäfers, O. Schober, and K. P. Schäfers, "List mode–driven cardiac and respiratory gating in PET," *J. Nucl. Med.*, vol. 50, no. 5, pp. 674–681, May 2009.

[5] M. Teräs, T. Kokki, N. Durand-Schaefer, T. Noponen, M. Pietilä, J. Kiss, E. Hoppela, H. T. Sipilä, and J. Knuuti, "Dual-gated cardiac PET–clinical feasibility study," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 37, no. 3, pp. 505–516, Mar. 2010.

[6] A. L. Kesner, P. J. Schleyer, F. Büther, M. A. Walter, K. P. Schäfers, and P. J. Koo, "On transcending the impasse of respiratory motion correction applications in routine clinical imaging—A consideration of a fully automated data driven motion control framework," *EJNMMI Phys.*, vol. 1, no. 1, p. 8, 2014.

[7] C. Chan, R. Fulton, D. D. Feng, and S. Meikle, "Median non-local means filtering for low SNR image denoising: Application to PET with anatomical knowledge," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf.*, Oct. 2010, pp. 3613–3618.

[8] N. Joshi, S. Jain, and A. Agarwal, "An improved approach for denoising MRI using non local means filter," in *Proc. 2nd Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Oct. 2016, pp. 650–653.

[9] Y. Q. Wang, J. Guo, W. Chen, and W. Zhang, "Image denoising using modified Perona–Malik model based on directional Laplacian," *Signal Process.*, vol. 93, no. 9, pp. 2548–2558, Sep. 2013.

[10] J. Bai and X. C. Feng, "Fractional-order anisotropic diffusion for image denoising," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2492–2502, Oct. 2007.

[11] A. Danielyan, V. Katkovnik, and K. Egiazarian, "BM3D frames and variational image deblurring," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1715–1728, Apr. 2012.

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3D transform-domain collaborative filtering," *Proc. SPIE*, vol. 6812, Mar. 2008, Art. no. 681207.

[13] T. F. Chan, J. Shen, and L. Vese, "Variational PDE models in image processing," *Notices AMS*, vol. 50, no. 1, pp. 14–26, 2003.

[14] T. Barbu, "A novel variational PDE technique for image denoising," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2013, pp. 501–508.

[15] J. Sliž and J. Mikulka, "Advanced image segmentation methods using partial differential equations: A concise comparison," in *Proc. Prog. Electromagn. Res. Symp. (PIERS)*, Aug. 2016, pp. 1809–1812.

[16] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 491–503, Feb. 2017.

[17] A. J. Reader, G. Corda, A. Mehranian, C. da Costa-Luis, S. Ellis, and J. A. Schnabel, "Deep learning for PET image reconstruction," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 1, pp. 1–25, Jan. 2021.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[20] Z. Zhang, C. Duan, T. Lin, S. Zhou, Y. Wang, and X. Gao, "GVFOM: A novel external force for active contour based image segmentation," *Inf. Sci.*, vol. 506, pp. 1–18, Jan. 2020.

[21] L. Xiang, Y. Qiao, D. Nie, L. An, W. Lin, Q. Wang, and D. Shen, "Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI," *Neurocomputing*, vol. 267, pp. 406–416, Dec. 2017.

[22] J. Xu, E. Gong, J. Pauly, and G. Zaharchuk, "200x low-dose PET reconstruction using deep learning," 2017, *arXiv:1712.04119*. [Online]. Available: http://arxiv.org/abs/1712.04119

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

[24] K. Gong, J. Guan, C.-C. Liu, and J. Qi, "PET image denoising using a deep neural network through fine tuning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 153–161, Mar. 2019.

[25] B. Zhou, Y.-J. Tsai, and C. Liu, "Simultaneous denoising and motion estimation for low-dose gated PET using a Siamese adversarial network with gate-to-gate consistency learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 743–752.

[26] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1867–1888, 2020.

[27] K. Gong, C. Catana, J. Qi, and Q. Li, "PET image reconstruction using deep image prior," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1655–1665, Jul. 2019.

[28] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," 2020, *arXiv:2008.13751*. [Online]. Available: http://arxiv.org/abs/2008.13751

[29] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," 2018, *arXiv:1803.04189*. [Online]. Available: http://arxiv.org/abs/1803.04189

[30] J. Schaefferkoetter, J. Yan, C. Ortega, A. Sertic, E. Lechtman, Y. Eshet, U. Metser, and P. Veit-Haibach, "Convolutional neural networks for improving image quality with noisy PET data," *EJNMMI Res.*, vol. 10, no. 1, pp. 1–11, Dec. 2020.

[31] L. Zhou, J. D. Schaefferkoetter, I. W. K. Tham, G. Huang, and J. Yan, "Supervised learning with cyclegan for low-dose FDG PET image denoising," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101770.

[32] D. Wu, H. Ren, and Q. Li, "Self-supervised dynamic CT perfusion image denoising with deep neural networks," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 3, pp. 350–361, May 2021.

[33] V. Bettinardi, L. Presotto, E. Rapisarda, M. Picchio, L. Gianolli, and M. C. Gilardi, "Physical performance of the new hybrid PET/CT discovery-690," *Med. Phys.*, vol. 38, no. 10, pp. 5394–5411, 2011.

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. NIPS Workshop Autodiff*, 2017, pp. 1–4.

[35] J. Howard. (2018). *Fastai*. [Online]. Available: https://github.com/fastai/fastai

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[37] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 558–567.
[38] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *CoRR*, vol. abs/1611.05431, 2016. [Online]. Available: https://arxiv.org/abs/1611.05431
[39] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," 2017, *arXiv:1707.02937*.
[40] G. Liu, K. J. Shih, T.-C. Wang, F. A. Reda, K. Sapra, Z. Yu, A. Tao, and B. Catanzaro, "Partial convolution based padding," 2018, *arXiv:1811.11718*. [Online]. Available: http://arxiv.org/abs/1811.11718
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[42] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318
[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[44] L. An, P. Zhang, E. Adeli, Y. Wang, G. Ma, F. Shi, D. S. Lalush, W. Lin, and D. Shen, "Multi-level canonical correlation analysis for standard-dose PET image estimation," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3303–3315, Jul. 2016.
[45] S. Akramullah, *Digital Video Concepts, Methods, and Metrics: Quality, Compression, Performance, and Power Trade-Off Analysis*. New York, NY, USA: Apress, 2014.
[46] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[48] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1—Learning rate, batch size, momentum, and weight decay," 2018, *arXiv:1803.09820*. [Online]. Available: http://arxiv.org/abs/1803.09820
[49] I. Buvat and C. Riddell, "A bootstrap approach for analyzing the statistical properties of SPECT and PET images," in *Proc. IEEE Nucl. Sci. Symp. Conf. Rec.*, vol. 3, Nov. 2001, pp. 1419–1423.
[50] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," 2014, *arXiv:1412.6596*. [Online]. Available: http://arxiv.org/abs/1412.6596

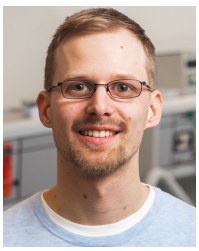
JOAQUIN RIVES GAMBIN was born in Dolores, Spain, in 1994. He received the B.Sc. degree in biotechnology, in 2016. He is currently pursuing the M.Sc. degree in medical analytics & health IoT with the University of Turku (UTU). He is also working as a Full-Time Researcher with the Health Technology Department, UTU. His research interests include biomedical signal and image processing, machine/deep learning, and development of software solutions for medical applications.

MOJTABA JAFARI TADI received the Ph.D. degree. His research at Stanford Molecular Imaging Program includes motion correction in PET/MRI, image denoising, bioinstrumentation, and machine learning. He is currently a Postdoctoral Research Fellow with the Stanford University School of Medicine. He is also working as a Principal Lecturer in artificial intelligence at Turku University of Applied Sciences (TUAS). His research interests include noninvasive physiological monitoring for human health, and data analysis, including developing signal processing and machine learning pipelines for detecting chronic diseases, computational medical imaging, and deep learning.

JARMO TEUHO received the M.Sc. degree in medical physics from the Tampere University of Technology, in 2012, and the Ph.D. degree in medical physics and engineering from the University of Turku, in 2018. He is currently positioned as a Researcher with the Turku PET Centre, Turku, Finland, where he has been working, since 2011.

RIKU KLÉN received the M.Sc. and Ph.D. degrees in mathematics from the University of Turku, Finland, in 2004 and 2009, respectively. He is currently an Assistant Professor with the Turku PET Centre, Turku, Finland. His research interests include medical imaging instrumentation and digital image processing.

JUHANI KNUUTI investigates on the physiology and pathophysiology, diagnosis and new therapies of coronary artery disease, heart failure, and metabolic diseases. The research is also focused on developing and utilizing novel noninvasive imaging methods (PET, SPECT, echocardiography, MRI, and CT) that will help to determine the risk and severity of CHD and heart failure and provide guidance for therapy decisions. The recent focus in cardiac research has been on vulnerable plaques, cardiac remodeling and multimodality and hybrid imaging. In metabolic and diabetes research, the focus has been in the interactions between different organs and heart in the pathogenesis and development of the cardiac diseases.



JUHO KOSKINEN received the Master of Science degree in technology from the University of Turku, Finland, in 2018. He has been working in health technology, since 2012. He is currently a Project Researcher with the University of Turku. His research interests include non-invasive heart motion monitoring and it's applications.

ANTTI SARASTE received the M.D., Ph.D., and FESC degrees. He is currently working as a Professor in cardiovascular medicine with the University of Turku and the Chief Physician with the Heart Center, Turku University Hospital, Turku, Finland.

EERO LEHTONEN (Member, IEEE) received the D.Sc. (Tech.) degree. He is currently working as a Senior Researcher with the Digital Health Technology Group, Department of Computing, University of Turku, Finland. He has been worked with the Health Technology Research, since 2012, and has supervised a doctoral thesis from mechanical cardiography and motion estimation in PET imaging. He has also experience from working as a Machine Vision Specialist in several companies. He is currently the PI with the Academy of Finland Research Project MINMOTION on applying MEMS technology to improve the quality of PET/CT.