Improved 3D Tumour Definition and Quantification of Radiotracer Uptake Using Deep Learning *†‡

Laura Dal Toso [1], Zacharias Chalampalakis[2], Irène Buvat[3], Claude Comtat[2], Gary Cook[1], Vicky Goh[1], Julia A. Schnabel[1], and Paul K. Marsden[1]

[1] *School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK*

[2] *Laboratoire d'Imagerie Biomédicale Multimodale (BioMaps), Université Paris-Saclay, CEA, CNRS, Inserm, Service Hospitalier Frédéric Joliot, Orsay, France*

[3] *Laboratoire d'Imagerie Translationnelle en Oncologie, Inserm, Institut Curie, Orsay, France*

Abstract

In clinical positron emission tomography (PET) imaging, quantification of radiotracer uptake in tumours is often performed using semi-quantitative measurements such as the standardised uptake value (SUV). For small objects, the accuracy of SUV estimates is limited by the noise properties of PET images and the partial volume effect. There is need for methods that provide more accurate and reproducible quantification of radiotracer uptake. In this work, we present a deep learning approach with the aim of improving quantification of tumour radiotracer uptake and tumour shape definition. A set of simulated tumours, assigned with "ground truth" radiotracer distributions, are used to generate realistic PET raw data which are then reconstructed into PET images. In this work, the ground truth images are generated by pasting simulated tumours, characterised by different sizes and activity distributions, in the left lung of an anthropomorphic phantom.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764458.

†This research was supported by the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], and by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

‡Cancer Research UK National Cancer Imaging Translational Accelerator Award (C4278/A27066)

These images are then used as input to an analytical simulator to simulate realistic raw PET data. The PET images reconstructed from the simulated raw data, and the corresponding ground truth images are used to train a 3D convolutional neural network. When tested on an unseen set of reconstructed PET phantom images, the network yields improved estimates of the corresponding ground truth. The same network is then applied to reconstructed PET data generated with different point spread functions. Overall the network is able to recover better defined tumour shapes and improved estimates of tumour maximum and median activities. Our results suggest that the proposed approach, trained on data simulated with one scanner geometry, has the potential to restore PET data acquired with different scanners.

1 Introduction

Positron emission tomography (PET) is an imaging modality which is extensively used in oncology to detect and stage tumours, and to monitor response to treatment. In clinical routine, image interpretation is usually performed by visual inspection of PET images, which leads to inter- and intra-observer variability. Although in many cases visual assessment may be sufficient, more challenging tasks such as the evaluation of the response to therapy of solid tumours require some form of quantification [1]. In PET, uptake quantification is usually performed using semi-quantitative measurements of tumour radiotracer uptake, such as the standardised uptake value (SUV). The SUV is defined as the ratio of the radioactivity concentration in a region of interest (ROI) to the average radioactivity concentration in the whole body and can be calculated as follows:

$$SUV = \frac{Activity\ Concentration_{ROI}\ (kBq/ml)}{Injected\ Activity\ (MBq)\ /\ Body\ Weight\ (kg)} \tag{1}$$

Usually a normalization by a mass density of 1 (g/ml) is assumed, and SUV in presented as a dimensionless metric. Different SUV metrics can be obtained, namely $\mathrm{SUV_{max}}$, $\mathrm{SUV_{mean}}$ and $\mathrm{SUV_{peak}}$. Most commonly $\mathrm{SUV_{max}}$, which is calculated using the most intense tumour voxel within the ROI, is reported. These semi-quantitative metrics are affected by the noise properties of PET images as well as partial volume effects (PVE), which compromise the accuracy and reproducibility of the measurements [1] [2]. The PVE, which results from the poor spatial resolution of PET scanners and from image sampling, degrades the quantitative accuracy of PET images and it can result in large biases on measures of tracer uptake in small tumours [3] [4]. $\mathrm{SUV_{max}}$ is widely used as it is user-independent, but it is also affected by noise. A metric that is less dependent on noise is $\mathrm{SUV_{mean}}$, but its disadvantage is that it depends on the delineation of the volume of interest (VOI) in which the measurement is performed. $\mathrm{SUV_{peak}}$ measures

the average activity concentration within a 1 cm3 spherical VOI. In literature different definitions of SUV_{peak} have been proposed, using different VOI shapes, sizes and locations [5]. In this paper, the 1 cm3 spherical VOI is located in the tumour region that yields maximum SUV_{peak}. SUV_{peak} is less affected by image noise than SUV_{max}, but it presents some issues when applied to small tumours, especially if they are smaller than the VOI in which the peak value is measured. There is need for reproducible methods that allow for more accurate quantification of tumour radiotracer uptake. Improved uptake quantification would lead to a more accurate assessment of response to treatment, especially at the early stages.

The use of artificial intelligence in the field of medical imaging has increased dramatically over the last decade. In PET imaging, machine learning and deep learning methods have been successfully applied to tumour segmentation, classification, automatic detection and image reconstruction [6, 7, 8, 9]. In recent years, deep learning methods have been used to denoise static PET images, and they have demonstrated better performance than traditional denoising approaches for various tracers and tasks [131-143]. The two main deep learning architectures that have been used for denoising are convolutional neural networks (CNNs) [10] and generative adversarial networks (GANs) [11]. In previous work we developed a deep learning algorithm, using a 3D CNN, with the aim to improve quantification of tumour radiotracer uptake in simulated PET images [12]. The network was trained on simulated "ground truth" images that presented 3D shapes with typical tumour activity distributions found in clinical FDG images and a corresponding set of simulated PET images. The network was able to robustly estimate the original activity, yielding improved images in terms of shape, activity distribution and quantification of activity. The main limitation of our previous work was that the PET images were simulated in a simplistic way, which did not take into account many of the effects that degrade the image quality in PET images.

Supervised deep learning methods require large amounts of labelled data, which are hard to obtain in PET imaging. Usually PET studies only comprise a relatively small number of patients. Furthermore, the true radionuclide distribution, which would correspond to the "label" of the PET images, is very difficult to obtain and rarely known. This limitation affects not only deep learning-based methods, but all the PET data processing methods (i.e. image reconstruction) which can never be fully evaluated in vivo. The use of phantoms and realistic PET simulators partially overcomes the lack of large labelled datasets. Monte Carlo simulation is the most commonly used technique to generate realistic PET data, but it has the disadvantage of being computationally very demanding. A number of analytical simulators have been developed to generate simulated PET images in a shorter time [13, 14]. While analytical simulators are not as accurate as Monte Carlo based simulators, they enable fast generation of PET data with realistic

noise properties, which makes them particularly useful for creating large numbers of datasets. The simulation of realistic tumours also has some limitations. Real tumours are often characterised by inter- and intra- tumour heterogeneity, and by complex spatial structures. The mathematical and computational models of cancer that have been implemented so far mainly focus on describing a few specific aspects of the disease [15]. These models are not able to capture all the characteristics of tumour biology. In clinical practice, biopsies are performed to obtain ground truth information on tumour composition, but this method only provides limited information as the samples are extracted from a small region, and they cannot provide a description of the whole tumour. In this work, we generated a dataset of tumours assigned with three different activity patterns. The simulated dataset offers a wide variety of tumour activity values and tumour shapes, located at different positions within the left lung of an anthropomorphic phantom. This dataset is then used to train and test a 3D convolutional neural network (CNN) that can recover the real activity distribution from the signal seen on the PET image. We build on our previous deep learning approach described in [12] by significantly enhancing the simulation of PET images, using an analytical simulator developed by Stute at al. [16] and by testing the proposed approach on two different simulated datasets.

2 Material and methods

2.1 Overview of the method

An overview of the proposed deep learning approach in presented in Fig. 1. Tumours with different sizes, shapes and activity distributions were simulated and subsequently placed in the left lung of an anthropomorphic phantom [17]. Different activities were assigned to each organ to create the "ground truth" images. Additionally, the corresponding attenuation maps were created. These two were used as input to an analytical PET simulator to generate PET raw data, which were subsequently reconstructed to provide the simulated PET images, as is detailed in 2.2. The ground truth images and simulated PET images were used to train a 3D CNN, which will be described in section 2.4.

2.2 Simulation of ground truth images

At first, tumour-like 3D shapes with different volumes, spanning 0.01 to 200 ml, were simulated. This set of tumours was composed of three groups, each corresponding to a different activity pattern: tumours filled with uniform activity, tumours split into halves (each assigned with a different activity), and hollow tumours with background activity in the inner part, to mimic necrotic regions. The thickness of the external layer of hollow tumours was

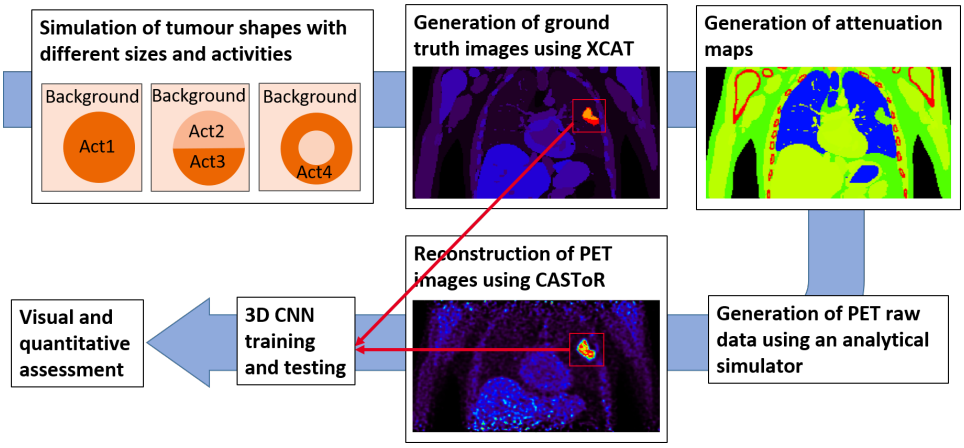


Figure 1: This figure shows an overview of the proposed method. Simulated tumours with various sizes and activities are simulated, and placed in random positions in the left lung of an anthropomorphic phantom, to generate the ground truth images. The attenuation maps are created, and used as input to an analytical simulator to generate the PET raw data. PET images are then reconstructed using CASToR [18]. 3D regions are cropped around the tumours and used to train a 3D CNN, which is then tested on an unseen set of reconstructed PET images.

set to half the radius of the tumour, and the inner core of the tumour was assigned with background activity. The tumour to background activity ratio was set to 1/10th. The choice of this specific set of activity patterns was based on the work of E. Pfaehler et al. [19], in which realistic phantom inserts were designed according to Non-Small Cell Lung Cancer (NSCLC) tumours extracted from patient studies. The simulated tumours were subsequently placed in random positions into the left lung region of the XCAT phantom [17]. In order to shorten the simulation times, the XCAT phantom was reduced to 344×344×127 voxels, to include only the torso. The voxel size was [2.09, 2.09, 2.03] mm^3. Realistic activity values, drawn from a range of radioactivity concentrations measured from real patients PET images, were assigned to each simulated tumour. Attenuation maps were generated for each XCAT image, using the XCAT phantom attenuation values.

2.3 PET data simulation and reconstruction

To generate realistic PET raw data we used an analytical PET simulator developed by Stute at al. [16]. The geometry, detector resolution, and sensitivity of a positron emission tomography/magnetic resonance (PET/MR) system, more specifically the Siemens mMR scanner (Siemens Biograph mMR, Erlangen, Germany), were used in the simulation. In this work we did not use any anatomical information from the MR, so only the PET component of

the simulator was modelled. PET raw data were generated from the ground truth images each with 100 million total prompt counts, including scatter and random events. This corresponds to 12.4 million Noise Equivalent Counts (NEC). Subsequently, image reconstruction was performed using the open source fully quantitative reconstruction platform CASToR [18], with an iterative ordered subset expectation maximization (OSEM) algorithm run to 6 iterations with 21 subsets. The reconstructed voxel size was [2.09, 2.09, 2.03] mm^3, as in the XCAT phantom images. Reconstructions were performed with image-based PSF modelling. Two different datasets were generated using the analytical simulator. The first one, called dataset 1, was composed of 800 uniform tumours, 721 tumours split in half and 589 hollow tumours. In this case the analytical simulator generated the PET data with an anisotropic, spatially invariant PSF with FWHM $(4.5, 4.5, 4.0)$ mm, which was also used in the reconstruction of the simulated PET images. A total of 2110 simulated PET raw datasets and images were generated.

One of the aims of this work, was to test if the proposed algorithm trained on a given dataset could generalise well to data acquired with different scanners. One way to simulate the diversity between different scanners is to generate and reconstruct PET images using a range of PSFs. As a result, a new dataset (called dataset 2) composed of 100 images, with 33 uniform tumours, 34 tumours in halves and 33 hollow tumours was generated using a range of anisotropic spatially-invariant PSFs. A value randomly drawn from a Gaussian distribution (mean μ= 4.5 mm and sigma σ= 0.2 mm) was assigned to the transaxial components of the PSF, both in the simulation and in the reconstruction, with a perfect match. The axial component was calculated by diving this value by 1.125, to maintain a constant ratio between transaxial and axial PSF components.

2.4 Network architecture

Convolutional neural networks (CNNs) are among the most commonly used algorithms for medical imaging applications [20]. These networks are composed of a series of convolutional layers, which can extract features from the input images, at multiple levels of abstraction. In this work, 3D CNNs with different depths were tested, and a visual and quantitative assessment suggested that a 3D CNN with 7 convolutional layers yielded the best results on our dataset. The proposed 3D CNN, presented in Fig. 2, is composed of 7 convolutional layers, each followed by a batch normalisation layer except for the final layer. The convolutional layers are characterised by 32 filters with dimensions 3×3×3, and by ReLU activation functions except for the final one which has linear activation function. Two dropout layers, with dropout rate 0.3, were added after the first and second batch normalisation layers. Mean squared error was used as loss function during training and the optimizer was Adam [21]. The learning rate was set to the default value

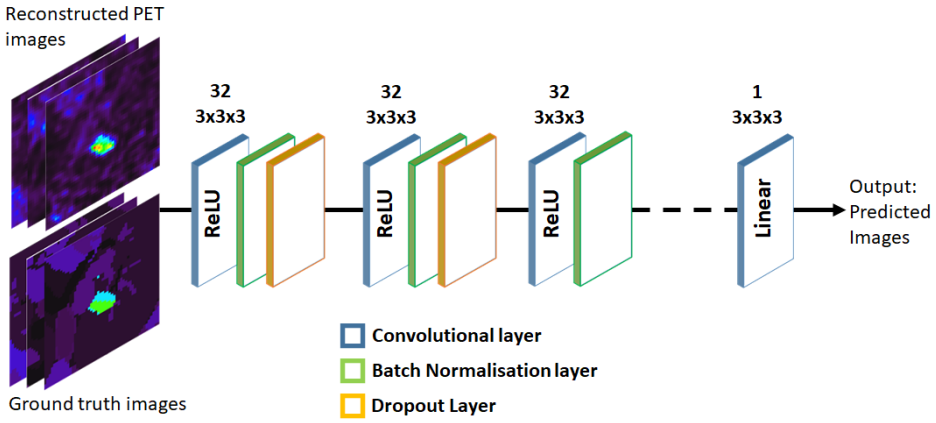


Figure 2: Reconstructed PET images and ground truth images, made of 50 × 50 × 50 voxels, were used as input to the 3D CNN for training. Once tested on an unseen set of reconstructed PET images the network yielded the corresponding predicted images.

0.001. The network was run on NVIDIA Tesla K40 GPU, and the network architecture was implemented in the Keras Framework with Tensorflow [22].

2.5 Experiments

All the simulated PET images were cropped around the tumours to a final dimension of 50×50×50 voxels, before being processed by the convolutional neural network. The network inputs were normalised using the MinMaxScaler, which is a function provided in the scikit-learn Python package [23]. Using this function, the training, testing and validation datasets were scaled in the range [0,1]. The normalisation factors were stored and subsequently applied to the network's predictions to rescale the resulting images, before performing any quantitative analysis. In all experiments, a visual assessment of the predicted images was at first performed using a free software tool for multimodality medical image analysis (AMIDE) [24]. Subsequently, the images were quantitatively assessed. In order to provide a baseline comparison, we used three metrics for the quantitative assessment of the results. The maximum, median and peak values were estimated for each tumour in order to quantitavely compare the reconstructed PET images to the images predicted by the CNN. In order to measure the maximum and median tumour uptake, tumour masks were defined on the ground truth images and subsequently applied to the reconstructed PET images and to the CNN's predicted images. Following the PERCIST guidelines [25], in this paper the peak value was measured within a 1 cm^3-volume spheric VOI, centred around the hottest voxel in the tumour mask. To present the results in a

more compact way, the recovery coefficients, defined as the ratio between the observed activity and the ground truth activity, were calculated using the maximum, median and peak values as shown in Eq. 2, Eq. 3 and Eq. 4 respectively.

$$RC_{max} = \frac{Max\ activity_{prediction}}{Max\ activity_{ground\ truth}} \tag{2}$$

$$RC_{median} = \frac{Median\ activity_{prediction}}{Median\ activity_{ground\ truth}} \tag{3}$$

$$RC_{peak} = \frac{Peak\ activity_{prediction}}{Peak\ activity_{ground\ truth}} \tag{4}$$

When testing the network on simulated data, the structural similarity index measure (SSIM) was also calculated. The structural similarity allows the comparison of two images by taking into account their luminance l, contrast c and structure s. The SSIM is defined as:

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{5}$$

where x and y are two non-negative image signals, which could be for example two image patches. α, β and γ are used to adjust the relative importance of the three components. In this paper, α, β and γ are equal to 1. The luminance comparison function l between two image signals depends on the mean intensities of the two signals μ_x and μ_y. After estimating the luminance, the mean intensity is removed from the initial image signal. At this point, contrast comparison between the two signals is performed by measuring their standard deviations σ_x and σ_y. Finally, each signal is normalised by its standard deviation and the structure comparison is made on the normalised signals. In practice, when two images (X,Y) are compared, the overall image quality can be estimated using the MSSIM, which it is expressed as:

$$MSSIM\,(X,Y) = \frac{1}{M}\sum_{j=1}^{M}(SSIM(x_j, y_j)) \tag{6}$$

where x_j and y_j are the image contents at the j-th local window and M is the number of local windows in the image. In this work, the MSSIM was used to compare the reconstructed PET images and the predicted images to the ground truth images.

2.5.1 CNN training and testing using PET data generated with a single PSF

This experiment was performed to optimise the network's architecture and test its performance on the simulated data. Dataset 1 was used to train

and test the network. The PET images were generated and reconstructed using a spatially invariant PSF with FWHM (4.5 mm, 4.5 mm, 4.0 mm). The simulated images were split into training and testing datasets, with ratio 80/20. 20 % of the training data are used for validation. The training dataset was augmented by scaling some of the bigger tumours with scaling factors ranging from 0.5 to 0.8, with the aim to increase the number of small tumours. As a result, the training dataset was composed of 645 uniform tumours, 645 tumours split in halves (of which 70 were augmented) and 645 hollow tumours (of which 177 were augmented). 20% of the images in the training dataset were used for validation. In this experiment, the network was trained for 500 epochs with batch size 50. The test dataset was made of 422 non-augmented images.

2.5.2 Application to PET data generated with different PSFs

The second aim of this work was to test if the proposed 3D CNN could generalise well to PET data generated with different PSFs, and restore these images accurately. More specifically, in this experiment the 3D CNN was trained on dataset 1, generated with one PSF as described in section 2.5.1, and subsequently applied to dataset 2, generated with a range of PSFs which were not learned during training.

3 Results

In this section, the results obtained training and testing the 3D CNN on data generated with a single PSF are presented. Subsequently, the same network is applied to PET data generated with a range of PSFs and the results are qualitatively and quantitatively compared to those obtained in the first experiment.

3.1 CNN training and testing using PET data generated with a single PSF

The first experiment was performed using the simulated training and test datasets generated with a single PSF, as described in section 2.5.1 . Three representative volumes, each characterised by a different activity pattern, are presented in Fig. 3. The ground truth images are shown in the first column, the corresponding reconstructed PET images are in the second column, and the CNN's predictions are shown in the last column. In this case, the CNN yielded better defined tumour shapes and denoised images. This visual assessment was followed by a quantitative analysis. In Fig. 4, 150 randomly selected MSSIM values are shown. The predicted images are overall characterised by a higher MSSIM to the ground truth images, but the absolute MSSIM values are below 0.65 in both datasets.

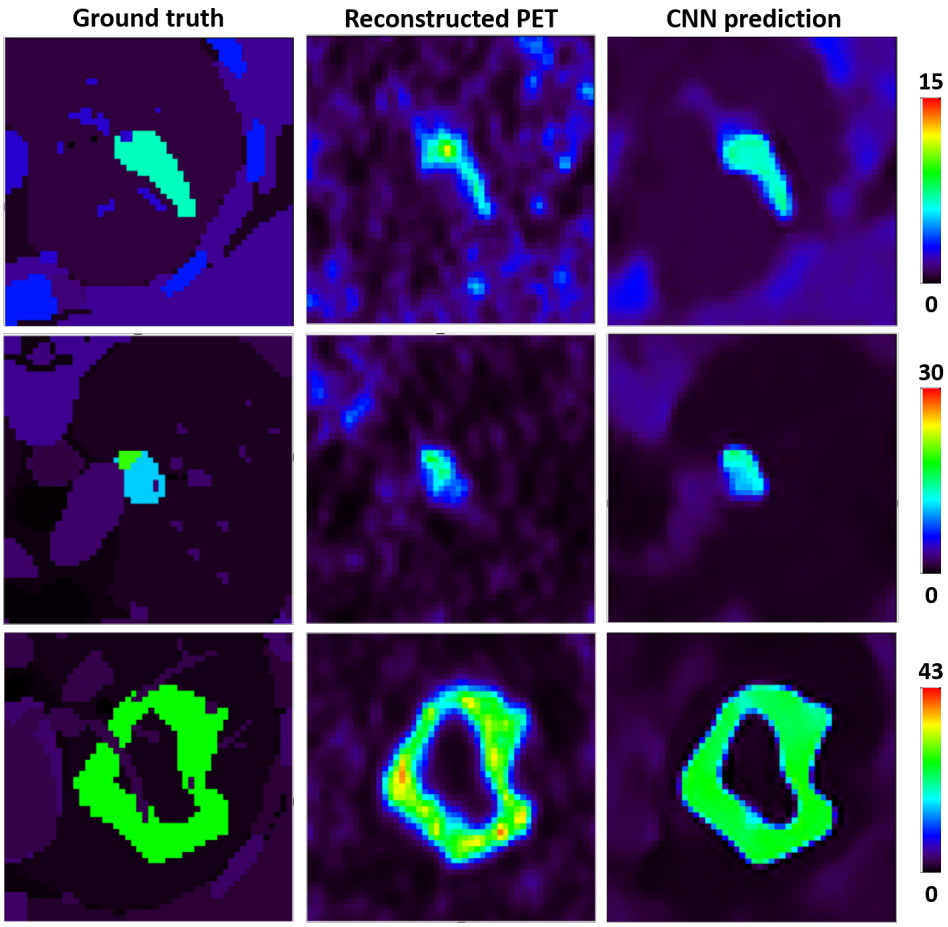


Figure 3: Transverse views of three representative volumes, belonging to experiment 1, where the reconstructed PET images were generated with one PSF. Each column shows the ground truth images, the reconstructed PET images and the CNN's predicted images respectively. In each row the images are shown with the same colour scale, expressed in kBq/ml.

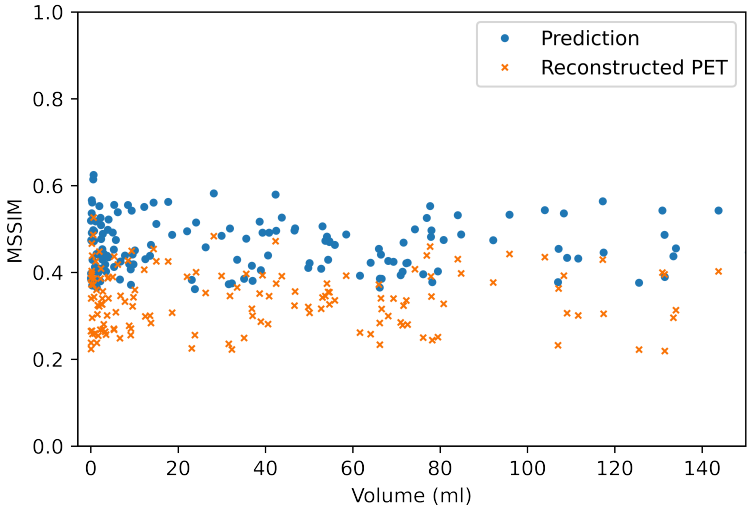


Figure 4: This figure shows 150 representative MSSIM values, obtained in experiment 1, where the reconstructed PET images were generated with one PSF. The MSSIMs between reconstructed PET images and ground truth images are shown in orange, the MSSIMs between predictions and ground truth are plotted in blue.

To further assess the performance of the network, the maximum, peak and median tumour uptakes were measured for each image and the recovery coefficients calculated. The RC_{peak} values, measured for the reconstructed PET images and for the predicted images, are shown in table 1. The RC_{peak} coefficients are close to 2 both for the CNN predicitons and for the reconstructed PET images. The CNN tends to underestimate SUV peak, especially when small tumours are included in the analysis, and it yields slightly improved RC_{peak} values when tumour volumes smaller than 5 ml are excluded from the analysis.

$\mathbf{RC_{peak}}$	all volumes	volume≥5 ml
Reconstructed PET	0.95±0.24	1.07±0.05
CNN's predictions	0.86±0.22	0.97±0.04

Table 1: Average RC_{peak} values relative to experiment 1, where the reconstructed PET images were generated with one PSF.

We further analysed the results by measuring the maximum and median tumour uptake. The median value was only estimated for the tumours with uniform uptake and for the hollow tumours, which also had uniform uptake. In Fig. 5 the RC_{max} and RC_{median} are plotted. Looking at Fig. 5a and

Fig. 5b it can be noted that the maximum and median values are not well recovered for tumours that have a volume smaller than around 5 ml. A detailed investigation of the small tumours was performed. In total, 5 volumes out of all the test images were predicted as false negatives, so the network predicted only background activity and did not recover a tumour volume. All of these tumours had a volume smaller than 0.18 ml (20 voxels). The maximum and median activities were not accurately recovered for tumours with volumes between 0.18 and 1.33 ml that were close to other structures (ribs, heart) and that were characterised by a low ground truth activity. The average RC_{max} and RC_{median} values, measured on the reconstructed PET images and on the CNN predictions, are summarised in table 2 and table 3 respectively. In both tables, the first column shows the RC averaged over the entire test dataset, whereas the second column shows the average over the tumours with volume larger than 5 ml. In both cases, the recovery coefficients measured for the predicted images are approaching 1, meaning that the CNN yields improved estimates of the maximum and median activity within the tumours.

$\mathbf{RC_{max}}$	all volumes	volume≥5 ml
Reconstructed PET	1.71±0.44	1.87±0.22
CNN's predictions	0.97±0.24	1.06±0.07

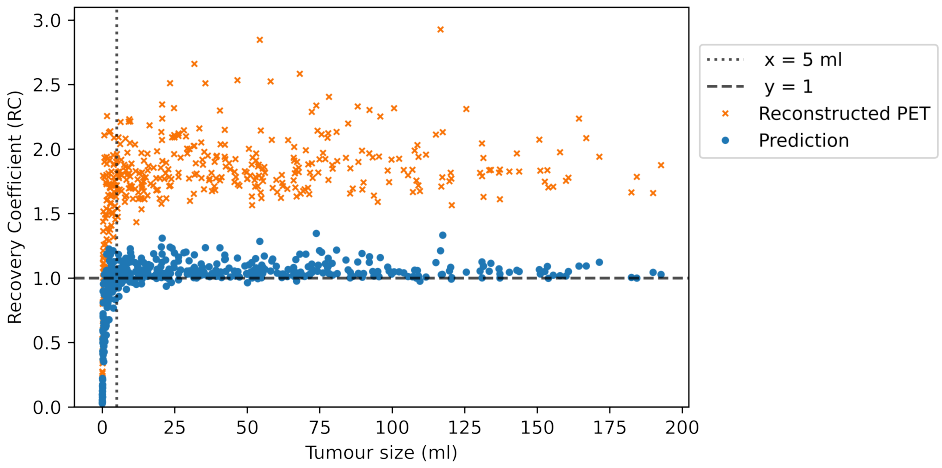
Table 2: Average RC_{max} values relative to experiment 1, where the reconstructed PET images were generated with one PSF. The values in the first column were calculated across the entire test set, the values in the second column were extracted only considering tumours larger than 5 ml.

$\mathbf{RC_{median}}$	all volumes	volume≥5 ml
Reconstructed PET	0.75±0.20	0.86±0.07
CNN's predictions	0.81±0.24	0.91±0.05

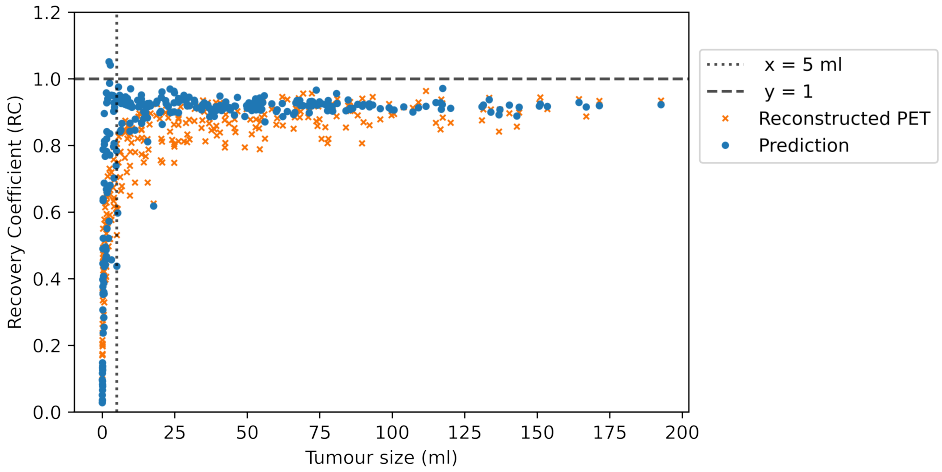
Table 3: Average RC_{median} values relative to experiment 1, where the reconstructed PET images were generated with one PSF. The values in the first column were calculated across the entire test set, the values in the second column were extracted only considering tumours larger than 5 ml.

3.2 Application to PET data generated with different PSFs

The network previously trained in experiment 1, where the reconstructed PET images were generated with one PSF, was then applied to PET data generated with different PSFs. In Fig. 6, three volumes belonging to the test dataset are presented. The PSFs used to simulate the PET raw data



(a) RC_{max}



(b) RC_{median}

Figure 5: The RC_{max} and RC_{median} values, obtained training and testing the network on PET data generated with a single PSF, are plotted against the tumour volume in (a) and (b) respectively. The coefficients measured using the reconstructed PET images are shown in orange, whereas the ones measured using the predicted images are shown in blue. Tumours split in halves, each assigned with a different activity, were excluded from the calculation of the median values.

and to reconstruct the PET images had FWHM (4.7, 4.7, 4.2) mm for the images in the first row, FWHM (4.0, 4.0, 3.6) mm for the images in the second row and FWHM (4.5, 4.5, 4.0) mm in the third row. Even though the PSFs used in the first and second row did not match the PSF used to generate and reconstruct the PET images used for training, the

$\mathbf{RC_{peak}}$	all volumes	volume≥5 ml
Reconstructed PET	0.95±0.23	1.07±0.08
CNN's predictions	0.86±0.21	0.95±0.05

Table 4: Average RC_{peak} values relative to PET data generated with different PSFs. The values in the first column were calculated across the entire test set, the values in the second column were extracted only considering tumours larger than 5 ml.

$\mathbf{RC_{max}}$	all volumes	volume≥5 ml
Reconstructed PET	1.78±0.46	1.94±0.26
CNN's predictions	0.99±0.24	1.07±0.08

Table 5: Average RC_{max} values relative to PET data generated with different PSFs. The values in the first column were calculated across the entire test set, the values in the second column were extracted only considering tumours larger than 5 ml.

network was still able to predict improved tumour shapes and activities. The RC_{peak} values, measured for the reconstructed PET images and for the CNN predictions are shown in 4. The results are comparable to the ones obtained in the previous experiment, the network only yields slightly improved RC_{peak} estimates when tumours smaller than 5 ml are excluded form the analysis. The RC_{max} and RC_{median} are plotted against tumour volume in Fig. 7. The median values were not calculated for the tumour split into two halves, each assigned with a different activity. The recovery curves show similar behaviours as in the previous experiment, and improved values are obtained when estimating the recovery coefficients on the images predicted by the CNNs. The average RC_{max} and RC_{median} measurements are presented in table 5 and table 6 respectively. The recovery coefficients calculated using the predicted images are comparable to the ones obtained in the previous experiment, proving that the network can successfully recover the maximum and median activity in the tumours even when tested on images generated with different PSF values.

4 Discussion and Conclusion

In this paper we propose a deep learning approach to improve quantification of radiotracer uptake and tumour shape definition in PET images. A 3D CNN was successfully trained and tested on simulated data generated with a single PSF, and applied to reconstructed PET images generated with a range of PSFs. The results indicate that the network is able to improve the definition of the tumour shapes and to denoise reconstructed PET images.

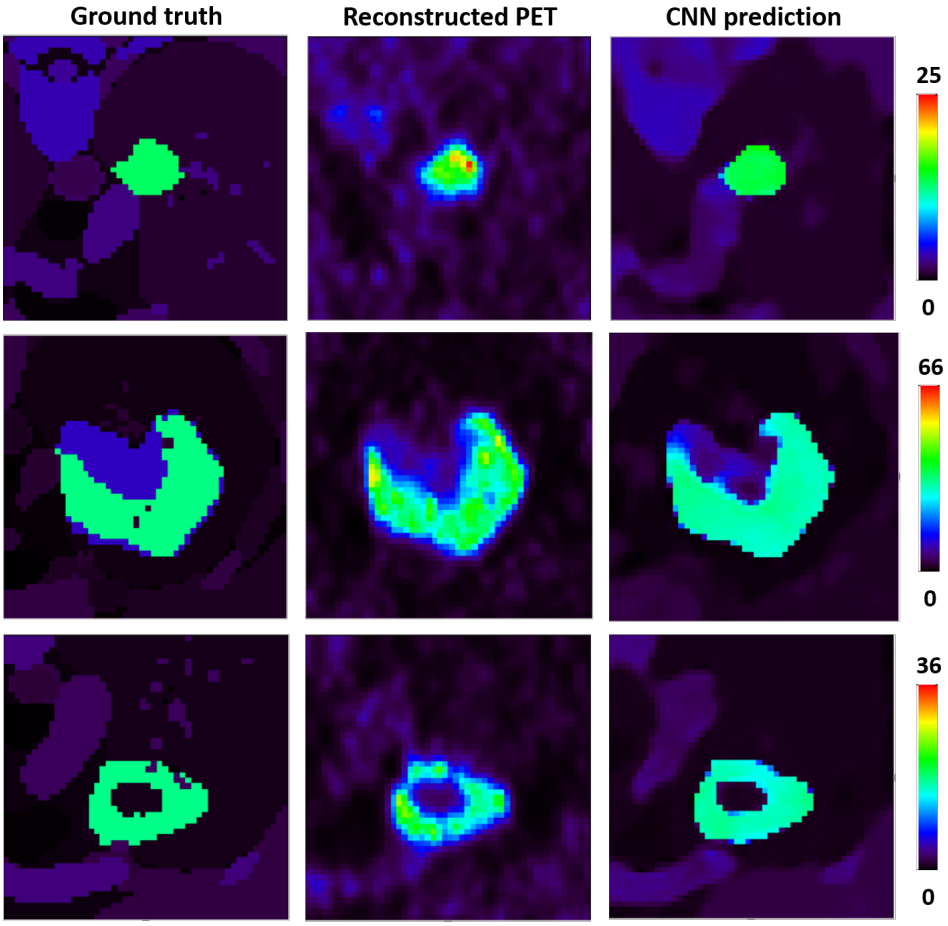
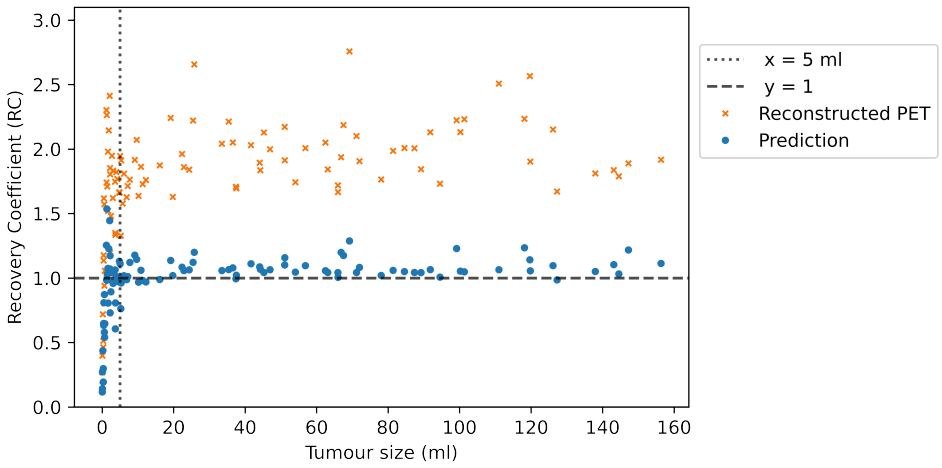


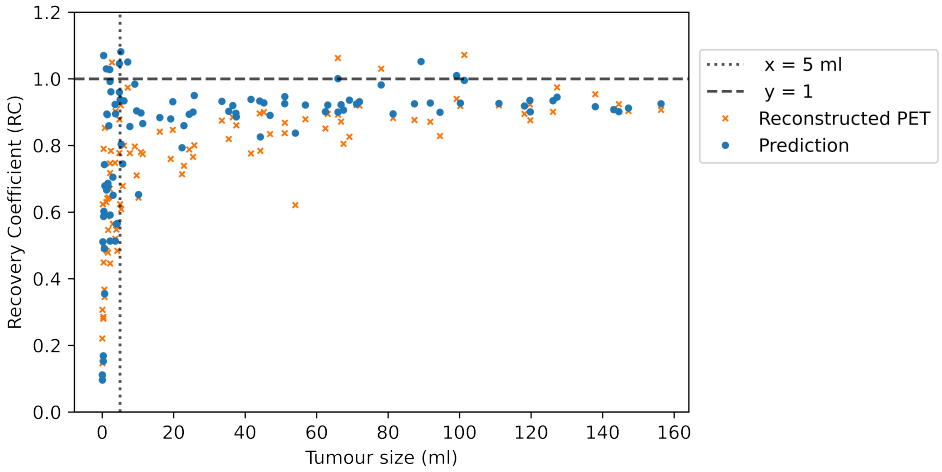
Figure 6: Transverse views of three representative volumes, generated each with a different PSF. The PET images were generated with a PSF with FWHM (4.7, 4.7, 4.2) mm in the first row, a PSF with FWHM (4.0, 4.0, 3.6) mm in the second row and PSF with FWHM (4.5, 4.5, 4.0) mm in the third row. Each column shows the ground truth images, the reconstructed PET images and the CNN's predicted images respectively. In each row the images are shown with the same colour scale, expressed in kBq/ml.

$\mathbf{RC_{median}}$	all volumes	volume≥5 ml
Reconstructed PET	0.75±0.18	0.84±0.08
CNN's predictions	0.83±0.20	0.90±0.05

Table 6: Average RC_{median} values relative to PET data generated with different PSFs. The values in the first column were calculated across the entire test set, the values in the second column were extracted only considering tumours larger than 5 ml.



(a) RC_{max}



(b) RC_{median}

Figure 7: The RC_{max} and RC_{median} values, obtained training and testing the network on PET data generated with a range of PSFs, are plotted against the tumour volume in (a) and (b) respectively. The coefficients measured using the reconstructed PET images are shown in orange, whereas the ones measured using the predicted images are shown in blue. Tumours split in halves, each assigned with a different activity, were excluded from the calculation of the median values.

A quantitative analysis of the results obtained using simulated data has shown that the images predicted by the 3D CNN yield improved estimates of the maximum tumour activities. We observed that the maximum and median activities were not accurately recovered for tumours with a volume smaller than 5 ml, so a more detailed analysis was performed on small tu-

mours. Only tumour volumes smaller than 0.18 ml presented critical issues, such as false negative predictions. Bigger volumes, which were close to other background structures and were characterised by a low ground truth activity, were generally associated with an inaccurate prediction of the maximum and median activity. In future work, we plan to further augment the training dataset, thus adding more small volumes to the training dataset. We think that this might improve the performance of the network for this class of tumours, as the network would be able to learn from more small tumours during training. A secondary effect that we noticed in our experiments was an improvement in the recovery of background structures. This effect will be further investigated in future work. Our approach proved successful when the network was applied to a set of reconstructed PET images generated with a range of PSFs that did not match the PSF used to generate the training dataset. These preliminary results suggest that the proposed approach would be able to restore PET images acquired with different scanners and spatially varying PSF. This work has the potential to be extended to larger areas of the body, in order to improve the estimation of the total tumour burden. The proposed approach has been tested on images generated with a single noise level, the robustness of this method to data generated with other noise levels remains to be evaluated. In this work, the same spatially invariant PSF was used for the simulation of the PET raw data and for the reconstruction of PET images. Further experiments are needed to test the proposed method using PET data generated with spatially invariant and non stationary PSFs, that would reproduce more closely clinical PET data. Furthermore, the PSFs used in this work to generate the PET raw data are characterised by the same ratio between axial and transaxial components, further experiments will be performed to assess the robustness of our deep learning approach to data generated using PSFs with a different anisotropy. In a clinical setting there may be a mismatch between the system's PSF and the one modelled in the reconstruction algorithm, this aspect will be investigated in future work. Finally, the proposed approach could be further tested including tumour shapes associated with more complex heterogeneous activities.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764458. This research was supported by the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], and by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] R. Boellaard, N. C. Krak, O. S. Hoekstra, and A. A. Lammertsma, "Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: A simulation study," *Journal of Nuclear Medicine* **45** no. 9, (2004) 1519–1527.

[2] H. Zaidi and N. Karakatsanis, "Nuclear medicin: physics special feature review article towards enhanced PET quantification in clinical oncology," tech. rep., 2017.

[3] M. Soret, S. Bacharach, and I. Buvat, "Partial-volume effect in PET tumor imaging," *Journal of Nuclear Medicine* **48** no. 6, (2007) 932–945.

[4] M. Cysouw, G. Kramer, L. Schoonmade, R. Boellaard, H. De Vet, and O. Hoekstra, "Impact of partial-volume correction in oncological PET studies: a systematic review and meta-analysis," *European Journal of Nuclear Medicine and Molecular Imaging* **44** (2017) 2105–2116.

[5] M. Vanderhoek, S. B. Perlman, and R. Jeraj, "Impact of the definition of peak standardized uptake value on quantification of treatment response," *Journal of Nuclear Medicine* **53** no. 1, (2012) 4–11.

[6] K. Gong, E. Berg, S. R. Cherry, and J. Qi, "Machine learning in PET: From photon detection to quantitative image reconstruction," *Proceedings of the IEEE* **108** no. 1, (2020) 51–68.

[7] K. Kim, D. Wu, K. Gong, J. Dutta, J. H. Kim, Y. D. Son, H. K. Kim, G. El Fakhri, and Q. Li, "Penalized PET reconstruction using deep learning prior and local linear fitting," *IEEE Transactions on Medical Imaging* **37** no. 6, (2018) 1478–1487.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I.

Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis* **42** (2017) 60–88.

[9] L. K. Shiyam Sundar, O. Muzik, I. Buvat, L. Bidaut, and T. Beyer, "Potentials and caveats of ai in hybrid imaging," *Methods* **188** (2021) 4–19. Artificial Intelligence Approaches for Imaging Biomarkers.

[10] K. Gong, J. Guan, C. C. Liu, and J. Qi, "PET image denoising using a deep neural network through fine tuning," *IEEE Transactions on Radiation and Plasma Medical Sciences* **3** no. 2, (2019) 153–161.

[11] Y. Wang, B. Yu, L. Wang, C. Zu, D. 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* **174** (2018) 550–562.

[12] L. Dal Toso, E. Pfaehler, R. Boellaard, J. A. Schnabel, and P. K. Marsden, "Deep learning based approach to quantification of PET tracer uptake in small tumors," in *Machine Learning for Medical Image Reconstruction*, F. Knoll, A. Maier, D. Rueckert, and J. C. Ye, eds., pp. 181–192. Springer International Publishing, Cham, 2019.

[13] B. Berthon, I. Häggström, A. Apte, B. J. Beattie, A. S. Kirov, J. L. Humm, C. Marshall, E. Spezi, A. Larsson, and C. R. Schmidtlein, "PETSTEP: Generation of synthetic PET lesions for fast evaluation of segmentation methods," *Physica Medica* **31** no. 8, (2015) 969–980.

[14] E. Pfaehler, J. Jong, R. Dierckx, F. van Velden, and R. Boellaard, "SMART (SiMulAtion and ReconsTruction) PET: an efficient PET simulation-reconstruction tool," *EJNMMI Physics* **5** (2018) .

[15] S. Bekisz and L. Geris, "Cancer modeling: From mechanistic to data-driven approaches, and from fundamental insights to clinical applications," *Journal of Computational Science* (2020) 101198.

[16] S. Stute, C. Tauber, C. Leroy, M. Bottlaender, V. Brulon, and C. Comtat, "Analytical simulations of dynamic pet scans with realistic count rates properties," pp. 1–3. 2015.

[17] W. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. Tsui, "4D XCAT phantom for multimodality imaging research," *Medical physics* **37** (2010) 4902–15.

[18] T. Merlin, S. Stute, D. Benoit, J. Bert, T. Carlier, C. Comtat, M. Filipovic, F. Lamare, and D. Visvikis, "CASToR: A generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction," *Physics in Medicine and Biology* **63** (2018) .

[19] E. Pfaehler, R. Beukinga, J. Jong, R. Slart, C. Slump, R. Dierckx, and R. Boellaard, "Repeatability of 18 F-FDG PET radiomic features: a phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method," *Medical Physics* **46** (2018) 665–678.

[20] R. Yamashita, M. Nishio, R. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging* **9** (2018) .

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations* (2014) .

[22] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, and X. Zheng, "TensorFlow : large-scale machine learning on heterogeneous distributed systems," 2015.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[24] A. Loening and S. Sam Gambhir, "AMIDE: a free software tool for multimodality medical image analysis.," *Molecular Imaging* **2** (2003) 131–137.

[25] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge, "From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors," *Journal of Nuclear Medicine* **50** no. Suppl 1, (2009) 1–50.