NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

LUÍSA CASTELBRANCO DA SILVEIRA COELHO SILVA

Bachelor of Science in Biomedical Engineering

DEVELOPMENT OF DEEP-LEARNING-BASED DENOISING ALGORITHMS FOR FAST WHOLE-BODY [^{18}F]FDG PET/CT SCANS

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
September, 2023

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

DEVELOPMENT OF DEEP-LEARNING-BASED DENOISING ALGORITHMS FOR FAST WHOLE-BODY [^{18}F]FDG PET/CT SCANS

LUÍSA CASTELBRANCO DA SILVEIRA COELHO SILVA

Bachelor of Science in Biomedical Engineering

Adviser: Dr. Francisco Paulo Marques de Oliveira
Researcher, Champalimaud Foundation

Co-adviser: Dr. Ricardo Nuno Pereira Verga e Afonso Vigário
Associate Professor, NOVA University Lisbon

Examination Committee

Chair: Dr. Hugo Filipe Silveira Gamboa
Associate Professor with Aggregation, NOVA University Lisbon

Members: Dr. Francisco Paulo Marques de Oliveira
Researcher, Champalimaud Foundation

Dr. Francisco José Santiago Fernandes Amado Caramelo
Assistant Professor, Coimbra University

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
September, 2023

Development of deep-learning-based denoising algorithms for fast whole-body [^{18}F]FDG PET/CT scans

Copyright © Luísa CastelBranco da Silveira Coelho Silva, NOVA School of Science and Technology, NOVA University Lisbon.
The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

Acknowledgements

I would like to set this page aside to thank everyone who contributed not only to this project, but to my academic journey over the last five years, both scientifically and/or personally.

First and foremost, I am extremely grateful to the Champalimaud Foundation, as its welcoming environment and its pedagogic policies provided me an experience I couldn't put into words if I tried. In the same line, I would like to express my deepest gratitude to Dr. Durval C. Costa, for having me amongst his brilliant team to develop my master's thesis. This opportunity will undoubtedly accompany me for the rest of my scientific career.

It is difficult to convey in a paragraph the never-ending patience, guidance, amity (and laughs) that filled the last few months, granted by my adviser, Dr. Francisco P. M. Oliveira. It is also difficult to summarise in a 40-page thesis everything I learned from him, not only about deep-learning-based denoising algorithms for fast whole-body [^{18}F]FDG PET/CT scans, but also about work-ethic, mentorship and kindness. Likewise, I would like to thank researcher Cláudia S. Constantino (the third, unofficial, co-adviser) for the approachability, the friendship, the continuous support and all the great advice (either thesis-related or not). Both have made it almost impossible to find an even remotely equal team.

I would also like to share the special appreciation I have for my colleagues from Champalimaud Research's room 19.10. The good environment and everyone's joyful disposition kept me motivated and made my days better.

Additionally, this study wouldn't have been feasible without the invaluable help of the Nuclear Medicine technicians of the Champalimaud Clinical Centre.

It is needless to say that it would not have been possible to come this far if not for the outstanding education I received at NOVA School of Science and Technology. I hold the highest esteem for my alma mater and for everyone that contributes to its remarkable excellence and its strikingly-hospitable environment. I am thankful to all the Professors that spread the passion for Biomedical Engineering that I carry with me, notably my co-adviser, Dr. Ricardo Vigário.

On a more personal note, I would like to thank my friends, without whom I wouldn't be able to function. Neves, Lopes, Afonso, Mary, Cate e Carol, I am grateful for you. Rosa e Braga, thank you for being the right examples, either academically or not. I would also like to thank my parents, for supporting my education and for instilling in me a restless curiousness and the love for learning. I would also like to thank my sisters, for giving me someone to look up to and for teaching me how to grow up. Last but not least, I would like to thank my grandfather, for the example of vocation and for showing me how to be proud of myself.

Abstract

This study aims to assess the feasibility of reducing the acquisition time of whole-body ^{18}F-labelled fluorodeoxyglucose ([^{18}F]FDG) positron emission tomography coupled with computed tomography (PET/CT) scans through deep-learning-based denoising.

112 whole-body [^{18}F]FDG PET/CT scans of patients with cancer were included. 92 were employed in the training of three convolutional neural networks: 2D, 2.5D and 3D U-Nets. Mean squared error (MSE) was the appointed loss function. The remaining 20 scans were set aside for testing. The images were acquired on a Philips Vereos Digital PET/CT scanner. From the standard-duration (70 seconds per axial field of view (AFOV)) raw data, fast scans were simulated by cropping the data to 15, 20 and 30 s/AFOV. Reconstruction was performed on-site using the manufacturer's protocol and following EARL1 standards. MSE, structural similarity index measure (SSIM) and intraclass correlation coefficient (ICC) were used for a voxel-wise comparison between the deep-learning-denoised (DL-denoised) fast scans and the reference images (70 s/AFOV). Signal-to-noise ratio (SNR) was computed in regions with expected uptake uniformity (liver and lungs) through the quotient between the mean standardised uptake value (SUV) and the SUV standard deviation. On a tumour basis, quantification was performed in terms of maximum SUV, mean SUV, SUV standard deviation, peak SUV, total lesion glycolysis (TLG) and metabolic tumour volume (MTV) in both the lesions in the DL-denoised and reference images. For benchmarking, Gaussian filter (GF), the state-of-the-art denoising method, was implemented and its width optimised in the training set through MSE minimisation relatively to the reference images.

The voxel-wise results revealed a strong agreement between the DL-denoised 15, 20 and 30-s/AFOV-based sets and the reference images, with an ICC equal or higher than 0.985. Quantification in the liver and lungs unveiled the DL-denoised images to have higher SNR compared to the original (fast), the GF-denoised and even the reference images. Tumour quantification exposed variations in the lesions' features that are not expected to have clinical impact, particularly in the 20 and 30-s/AFOV-based sets. Deep-learning-based denoising outperformed optimised Gaussian filter in every instance.

The deep-learning-based denoising models for fast whole-body [^{18}F]FDG PET/CT scans developed in this study proved to have potential to achieve images with clinically-suitable quantitative parameters. The 20 s/AFOV scans with post-processing with the 2.5D U-Net or the 3D U-Net seemed to be the best compromise between scan duration and image quality, compared to the 15 and 30-s/AFOV-based scans.

Keywords: [^{18}F]FDG PET/CT, deep learning, denoising, molecular imaging, oncology

Resumo

O objetivo deste estudo é avaliar a exequibilidade da redução do tempo de aquisição de exames de tomografia por emissão de positrões aliada a tomografia computadorizada (PET/CT) com fluorodesoxiglicose marcada com flúor-18 ([^{18}F]FDG) de corpo inteiro, através de um pós-processamento com aprendizagem profunda (*DL*, do inglês *deep learning*).

112 exames PET/CT com [^{18}F]FDG de corpo inteiro de pacientes com cancro foram incluídos. 92 foram utilizados no treino de três redes neuronais convolucionais: *U-Nets* 2D, 2.5D e 3D. O erro quadrático médio (*MSE*) foi utilizado como função-objetivo. Os restantes 20 exames foram reservados para teste. As imagens foram adquiridas num equipamento PET/CT Digital Philips Vereos. Das imagens padrão (70 segundos por campo de visão axial (*AFOV*)), imagens rápidas foram simuladas, cortando os dados para obter 15, 20 e 30 s/*AFOV*. A reconstrução foi feita localmente com o protocolo do fabricante e seguindo as normas EARL1. O *MSE*, a medida do índice de semelhança estrutural (*SSIM*) e o coeficiente de correlação intraclasse (*ICC*) foram utilizados para comparação vóxel-a-vóxel entre as imagens pós-processadas com *DL* e as de referência (70 s/*AFOV*). A razão sinal-ruído (*SNR*) foi calculada em regiões em que se espera captação uniforme (fígado e pulmões) através do quociente entre o valor médio de captação padrão (*SUV*) e o desvio padrão de *SUV*. Para a quantificação nos tumores, foram utilizados o *SUV* máximo, o *SUV* médio, o desvio padrão de *SUV*, o pico de *SUV*, a glicólise total da lesão (*TLG*) e o volume metabólico do tumor (*MTV*). Para análise comparativa, o filtro Gaussiano (*GF*) foi implementado e otimizado para o conjunto de treino por meio da minimização do *MSE* relativamente à referência.

Os resultados da análise vóxel-a-vóxel revelaram forte concordância entre as imagens pós-processadas com *DL* e as de referência, com um *ICC* igual ou superior a 0.985. A quantificação no fígado e nos pulmões mostrou que as imagens pós-processadas com *DL* apresentavam uma maior *SNR* do que as imagens rápidas originais, as pós-processadas com *GF* e mesmo do que as de referência. A quantificação nas lesões mostrou uma variação nas características que não se espera ter impacto clínico, em particular nas imagens resultantes das aquisições com 20 e 30 s/*AFOV*. O pós-processamento com *DL* superou o obtido com o *GF* em todas as ocasiões.

A remoção de ruído de imagens PET/CT com [^{18}F]FDG de corpo inteiro rápidas com base nos modelos de *deep learning*, desenvolvidos neste estudo, mostrou ter potencial para conseguir imagens com parâmetros quantitativos adequados para a clínica. As imagens adquiridas com 20 s/*AFOV* e pós-processamento, seja com a *U-Net* 2.5D seja com a 3D, pareceram ser o compromisso mais indicado entre a redução do tempo de aquisição e a qualidade de imagem.

Palavras-chave: [^{18}F]FDG PET/CT, aprendizagem profunda, remoção de ruído, imagem molecular, oncologia

Promotion and Dissemination

In this page, the different ways of dissemination which the work carried on in this study gave way to and will give way to are itemised.

International Conferences

- L. C. Silva, C. S. Constantino, M. Silva, F. P. M. Oliveira, R. Vigário, D. C. Costa. "Image quantitative parameters using deep learning-based denoising of ultra-fast whole-body [^{18}F]FDG PET/CT are comparable to standard acquisitions". EANM'23 Abstract Book Congress Sep 9-13, 2023. *European Journal of Nuclear Medicine and Molecular Imaging* (2023), p. S724. DOI: 10.1007/s00259-023-06333-x.

Articles

- L. C. Silva et al. "Deep-learning-based denoising of whole-body [^{18}F]FDG PET/CT allows ultra-fast acquisitions". *(in preparation)*

Grand-Challenge Participations

- L. C. Silva, C. S. Constantino, F. P. M. Oliveira, Durval C. Costa. "3D U-Net adaptation for fast-scan/low-dose whole-body [^{18}F]FDG PET/CT denoising". Ultra-low Dose PET (UDPET) Imaging Challenge 2023. URL: ultra-low-dose-pet.grand-challenge.org.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

[^{18}F]FDG	fluorine-18 fluorodeoxyglucose
2.5D	two-and-a-half-dimensional
2D	two-dimensional
3D	three-dimensional
AFOV	axial field-of-view
AI	artificial intelligence
ANN	artificial neural network
BMI	body mass index
BSREM	block sequential regularised expectation maximisation
CNN	convolutional neural network
CPU	central processing unit
CT	computed tomography
CuDNN	CUDA Deep Neural Network
DenseNet	dense convolutional neural network
DL	deep learning
EANM	European Association of Nuclear Medicine
EARL	EANM Research Ltd.
FWHM	full width at half maximum
GF	Gaussian filter
GPU	graphics processing unit
ICC	intraclass correlation coefficient
LOR	line of response

ML	machine learning
MSE	mean squared error
MTV	metabolic tumour volume
OSEM	ordered subset expectation maximisation
PET	positron emission tomography
PSMA	prostate-specific membrane antigen
ReLU	rectified linear unit
ResNet	residual convolutional neural network
ROI	region of interest
SNR	signal-to-noise ratio
SSIM	structural similarity index measure
SUV	standardised uptake value
$\mathbf{SUV_{max}}$	maximum standardised uptake value
$\mathbf{SUV_{mean}}$	mean standardised uptake value
$\mathbf{SUV_{peak}}$	peak standardised uptake value
$\mathbf{SUV_{SD}}$	standardised uptake value standard deviation
TLG	total lesion glycolysis

1 | Introduction

The research work described in this dissertation was carried out in accordance with the norms established in the ethics code of NOVA University Lisbon. The work described and the material presented in this dissertation, with the exceptions clearly indicated, constitute original work carried out by the author.

1.1 Context and Motivation

Cancer incidence and mortality are rapidly increasing worldwide, with cancer being, for most developed countries, the leading cause of premature death [1]. As a result, concerns about the prevention, diagnosis and treatment of cancer fuel global efforts for research and improvement of technological and clinical techniques in this area.

Positron emission tomography (PET), coupled with computed tomography (CT), is a key medical imaging modality in nuclear medicine. Ever since its first implementation for clinical evaluation in the 1990s, PET/CT has been a growing imaging technique in oncology [2] and neurology [3], as well as in other medical specialities. Referral to a PET/CT scan, in oncology, usually arises if the clinician suspects the presence of active tumours. In neurology, it may be performed to detect some neurological disorders such as Alzheimer's or Parkinson's disease.

PET scans require patients to be still for the duration of the exam, which is often painful and uneasy, and increases the susceptibility of motion artefacts during the acquisition. Other concerns raised are the exposure to radiation of both the patient and the technicians, as well as the regard for energy efficiency and rentability of the acquisition system. There is, therefore, a common interest to optimise this technique in terms of the patient's comfort, radiological protection and sustainability.

In recent years, artificial intelligence (AI) and, particularly, machine learning (ML), have been found to have numerous applications in healthcare [4]. Let it be for data handling or software performance and innovation, AI is a propitious tool that constitutes the core of many recent biomedical research methods. Deep learning (DL) is an ML technique that allows a system to build and learn concepts, with an increasing level of complexity and abstraction that attempts to mimic the human brain. Apropos of PET/CT, it has been used to simulate, reconstruct or enhance images, with promising clinical applications [5].

In this setting, the motivation of this thesis is to explore the potential of deep learning in whole-body fluorine-18 fluorodeoxyglucose ([^{18}F]FDG) PET/CT image processing, and

reinforce these methods in their clinical relevance and reproducibility.

1.2 Objectives

This study's main aim is to achieve a reliable deep-learning-based denoising algorithm for fast whole-body [^{18}F]FDG PET/CT scans, that outperforms standard post-processing methods and allows for a reduction of PET acquisition duration. Thus, the goal is to improve PET imaging in efficiency and in terms of the patient's comfort. This broad aim can be divided into three more specific objectives:

1. To develop a deep-learning-based algorithm to denoise fast [^{18}F]FDG PET/CT scans, intending to restore them to their standard quality.
2. To compare different artificial neural network architectures.
3. To compare the developed model against standard post-processing methods.

1.3 Literature Review

Given their already mentioned relevance, some DL-based algorithms have already been implemented and commercialised with the aim of PET/CT image denoising. Consequently, some studies focus solely on testing and validating one of the available softwares [6]–[9], while others develop and train their models [10]–[12].

SubtlePET™ is an already commercialised software that follows the architecture of a two-and-a-half-dimensional (2.5D) U-Net convolutional neural network (CNN). A significant amount of the publications found in the literature, concerning the denoising of PET images through deep learning, does a performance assessment of SubtlePET™ on different datasets. Bonardel et al. [7] aimed to test the limits of denoising statistically reduced PET raw data from 3 different PET scanners with SubtlePET™, in comparison to the standard-quality images of phantoms and patients. The simulated whole-body [^{18}F]FDG PET/CT images with ½ and ⅓ of the standard acquisition duration (of either 90 or 120 seconds per bed position) from 110 patients and 3 phantoms were post-processed with SubtlePET™ and subsequently evaluated. The results showed that, for both the phantom and patient datasets, the combination of the ½-acquisition-time fast-scan PET with SubtlePET™ post-processing, allowed similar qualitative (noise level) and quantitative (maximum standardised uptake value (SUV_{max})) parameters, when compared to the standard quality PET images. Weyts et al. [6] used the data from 195 patients referred for whole-body [^{18}F]FDG PET/CT to compare the reference acquisition (90 seconds per bed position) with the simulated ½-duration scans, with and without post-processing with SubtlePET™. The findings indicated that SubtlePET™ allows quality restoration of the half-acquisition-time scans.

Hyper DLR is a denoising software based on a 2.5D U-Net CNN that includes residual and dense connections as implemented on residual convolutional neural networks (ResNets) and

dense convolutional neural networks (DenseNets), respectively. Xing et al. [10] employed 313 whole-body [^{18}F]FDG PET/CT studies for training this network and 80 for validation. The images were acquired on a uMI 780, United Imaging Healthcare, digital PET/CT scanner. Network implementation was done using the *PyTorch* framework on Python 3.7, resorting, then, to the NVIDIA CUDA Deep Neural Network (CuDNN) library. The training pairs consisted of the simulated ½-duration scans as the noisy input and full-duration scans (180 seconds per bed position) as the reference. For testing purposes, 52 whole-body PET scans of full, ¾, ½ and ⅓-acquisition-time were generated, and post-processed both with Hyper DLR and Gaussian filter. The qualitative and semi-quantitative analysis of the resulting images revealed an improvement in the noise reduction ability for Hyper DLR denoising, when compared with Gaussian filter denoising.

Mehranian et al. [11] developed a three-dimensional (3D) U-Net to train three models to restore the simulated ¾, ¼ and ½-duration [^{18}F]FDG PET/CT scans to their full-duration (147±8 seconds per bed position) quality. The network was implemented with *PyTorch*. The images included in the dataset were acquired with either the GE Discovery MI or the GE Discovery 710 PET/CT scanners. The partial-time scans were reconstructed with the ordered subset expectation maximisation (OSEM) algorithm, and the full-time scans with both the OSEM and the block sequential regularised expectation maximisation (BSREM) algorithms. 277 examinations were split into 237 for training, 15 for validation and 25 for testing. A comparison through SUV_{max} values and a clinical evaluation by experienced radiologists showed that the post-processing with DL allows a scan time reduction of 25% versus full-time BSREM-reconstruction scans and of 50% versus full-time OSEM-reconstruction scans. Since OSEM is a faster reconstruction algorithm than BSREM, the reductions in scan time with OSEM reconstruction and post-processing with DL also reduced the total time (reconstruction plus denoising), as the result revealed to be equivalent to full-time BSREM-reconstruction scans.

Tsuchiya et al. [12] trained a deep residual convolutional neural network with 8 layers for the denoising of PET images with different levels of noise. The training set contained 6 lung and 2 brain [^{18}F]FDG PET/CT long-duration scans (14 and 15 minutes per bed) as the reference. These were used to simulate shorter-duration scans, with 30, 45, 60, 120, 180, 240, 300 and 420 seconds per bed position, to be employed as the network's noisy inputs. The test set comprised 50 whole-body [^{18}F]FDG PET/CT scans acquired with 120 seconds per bed position. The results showed that the deep-learning-based denoising method enhanced image quality without losing quantitative information, with better perceived image quality when compared with denoising through Gaussian smoothing.

This short review presents a summary of the progress done on PET/CT image denoising through deep learning so far, summing up the different approaches by different research teams. The designed studies roughly consist of reducing PET scans' data, to simulate low-activity/fast-scans, and use an artificial neural network as a denoising method. The resulting image is compared to the standard-quality full-duration or high-quality long-duration scans, post-processed with the ordinarily employed methods. All the reviewed studies exhibit promising results, concluding that deep learning outperforms most regular denoising methods. Table

1.1 presents a small overview of the reviewed studies.

Table 1.1: Overview of [^{18}F]FDG PET/CT deep-learning-based denoising strategies.

Study	Acquisition time	Network	Results
Bonardel et al. (2022) [7]	½ and ⅓ of the standard acquisition (90 or 120 s/AFOV)	SubtlePET™ (2.5D U-Net)	½-acquisition-time with DL post-processing presents similar qualitative and quantitative parameters when compared to the standard acquisition
Mehranian et al. (2022) [11]	¾, ½ and ¼ of the standard acquisition (147±8 s/AFOV)	3D U-Net	¾-acquisition-time with DL post-processing presents similar quality to standard-acquisition and reconstruction with the BSREM algorithm; ½-acquisition-time with DL post-processing presents similar quality to standard-acquisition and reconstruction with the OSEM algorithm
Tsuchiya et al. (2021) [12]	1/28, 3/56, 1/14, 1/7, 3/14, 2/7, 5/14 and ½ of the standard acquisition (14 or 15 min/AFOV)	2.5D ResNet	DL post-processing achieves better image quality than Gaussian smoothing without losing quantitative information
Weyts et al. (2022) [6]	½ of the standard acquisition (90 s/AFOV)	SubtlePET™ (2.5D U-Net)	½-acquisition-time with DL post-processing presents similar quality to the standard acquisition
Xing et al. (2022) [10]	¾, ½, ⅓ of the standard acquisition (180 s/AFOV)	Hyper DLR (2.5D U-Net with residual and dense connections)	DL post-processing outperforms Gaussian filtering

These studies constitute a solid baseline for the validation of the applications of deep-learning-based denoising methods for fast whole-body [^{18}F]FDG PET/CT scans. The present study, as outlined in the previous section, aims to further consolidate the feasibility of these methods and to explore new strategies.

2 Positron Emission Tomography

2.1 PET/CT

As introduced before, PET/CT is a non-invasive imaging technique, widely used in oncology and other fields. PET provides functional imaging by detecting the radioactive activity in the body, after a radiopharmaceutical is administered to the patient. This process is described in further detail in the following sections. CT provides structural imaging, as it measures X-ray attenuation by the different tissues in the body. Therefore, the combined efforts of PET and CT result in a hybrid image with both functional and anatomical information.

2.1.1 Radiopharmaceuticals

Radiopharmaceuticals are the foundation of PET imaging. These radioactive tracers are obtained by binding a radioisotope to a biological molecule that targets a given organ, tissue or function. The associated metabolic/functional activity is translated by the detection of the radioactive decay inherent to the bound radioisotope.

[^{18}F]FDG is employed to trace glucose consumption throughout the body, and it is administered by intravenous injection. Malignant tumorous cells have high metabolic activity and, thus, will need energy to grow and multiply. This way, [^{18}F]FDG will expectedly be consumed in larger amounts by the lesions, compared to the remaining tissues and structures. The [^{18}F]FDG PET scan will reveal these regions of glucose avidity in contrast to the low/normal activity ones.

2.1.2 Imaging Principle

Fluorine-18 decays through positron emission (β^+ decay), in which a proton of the radionuclide's nucleus is converted into a neutron. This involves the release of not only a positron (anti-electron, e^+) but also an electron neutrino (ν_e), as summarised by equation 2.1.

$$^{18}_{9}F \rightarrow ^{18}_{8}O + e^+ + \nu_e \tag{2.1}$$

When Fluorine-18 is present in the body, as it is after the administration of [^{18}F]FDG, the

positron emitted will almost immediately collide with an electron (e^-), leading to the emission of two γ rays, each with 511 keV of energy (eq. 2.2). Before its annihilation, the positron loses energy through interactions with the matter, reaching thermalisation.

$$e^+ + e^- \rightarrow \gamma + \gamma \tag{2.2}$$

These two photons will travel in approximately opposite directions. If they are stopped in the scanner's detectors, the event constitutes a so-called coincidence [13]. To a coincidence event is assigned a line of response (LOR) that intersects the two detectors. Each pair of photons that resulted from the same annihilation and are detected within a specific time-frame counts as a true coincidence. These constitute the actual PET signal that shows the spatial distribution of the positron-emitting decays. Other types of coincidences that result in a wrongful LOR constitute statistical noise in the resulting image. A scattered coincidence occurs when the photon suffers Compton scattering and is deviated from its original trajectory. A random coincidence occurs when two photons from different positron annihilation events are detected within the same coincidence time window of the system. The three types of coincidences are illustrated in fig. 2.1.

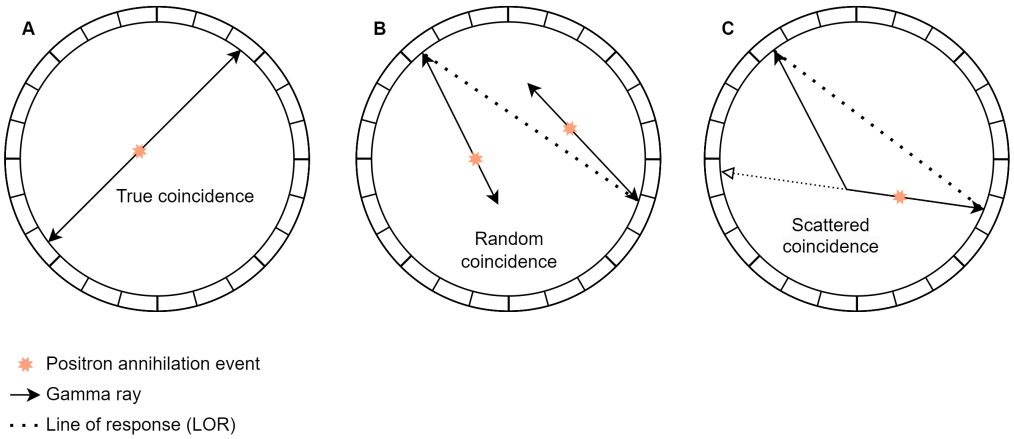


Figure 2.1: Types of coincidences in PET imaging. (A) True coincidence. (B) Random coincidence. (C) Scattered coincidence.

The ring covering the axial extent of the scanner, also illustrated in fig. 2.1, is composed of scintillation detectors. These crystal detectors are able to convert high-energy photons into visible light, through the interaction of the photons (through Compton scattering or photoelectric absorption) with the electrons of the material. These now-energetic electrons lose energy when interacting with the surrounding matter, thus exciting other electrons. In the process of returning to the ground state, the excited electrons emit light, which is converted to an electrical signal. This last step can be carried out through one of many available technologies. In the first PET scanners, photomultiplier tubes were used to this effect [13]. The Philips Vereos Digital PET/CT scanner, which was employed in this study, is equipped with digital silicon

photomultipliers, coupled with scintillator crystals, in a time-of-flight technology [14].

2.1.3 Data and image

The PET scan provides a distribution map of the registered radioactive activity in the body. This is translated into a volumetric image in which each voxel contains information about the corresponding activity concentration or standardised uptake value (SUV). SUV is a semi-quantitative measure of relative uptake, as the name suggests. It depends on the patient's weight, the concentration of radiopharmaceutical in each given voxel and the injected activity, as equation 2.3 reflects. If the radiopharmaceutical was distributed uniformly in the body, the SUV in each voxel would be 1 g/mL [15, Chapter 5].

$$\mathrm{SUV} = \frac{C_{\mathrm{PET}}}{\text{Injected activity}/\text{Weight}} \quad [\mathrm{g/mL}] \tag{2.3}$$

Where C_{PET} is the tissue's activity concentration in Bq/mL at the time of the acquisition.

In general, SUV values greater than 2.5 g/mL are considered abnormal [15, Chapter 11], if not in a region of expected high uptake.

2.2 Image Quality

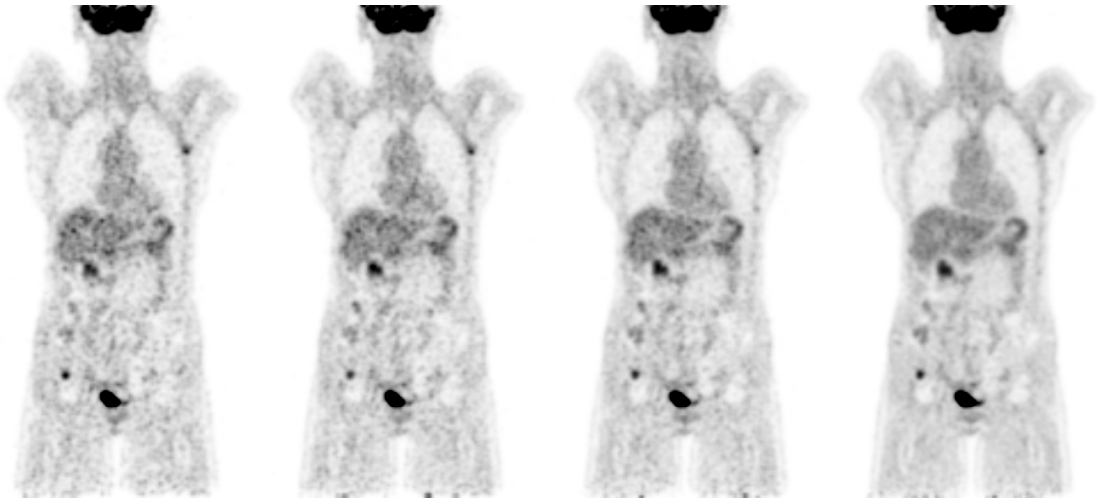
The resulting image must have enough quality for its clinical evaluation. Image corruption by noise or artefacts is inherent to medical imaging, occurring during acquisition and/or certain processing stages. The presence of noise is critical in small or low-contrast areas of the image, as it can lead to the loss of important clinical information. In PET imaging, image quality depends, to a large extent, on both the activity administered and the acquisition time. The limiting factor is, mainly, the administered activity, given the context of radiological protection of the patient and technicians. Relatively low activity is recommended in the guidelines for oncological [^{18}F]FDG PET imaging [16], as per the indications of the International Commission on Radiological Protection. This must be counterbalanced by a longer acquisition time, which, nevertheless, must be short enough to ensure the patient's comfort. However, the lower the activity and acquisition time, the noisier the ensuing image. It is, hence, relevant to study denoising methods for PET scans.

Despite having multiple sources, PET noise is considered to follow a mixed Gaussian-Poisson model [17]. The denoising process can be performed either during the reconstruction or in the form of post-processing. The former refers to denoising modules embedded in the reconstruction model that aim to reduce noise and enhance important information. The latter takes advantage of digital processing applied to the resulting image to reduce noise and improve quality. The most common post-processing technique for PET image denoising is Gaussian filtering or similar procedures. Some other approaches like adaptive filters have been looked into, as well as machine-learning-based methods [17].

2.2.1 Fast Scans

For the majority of PET scanners, the whole-body acquisition is performed sequentially for sets of axial slices along the body, scanned at each bed position. Depending on the axial field-of-view (AFOV), more or fewer bed positions will be needed to cover the whole body. In other words, the AFOV defines the extent of the body that will be scanned in a given bed position and, since it does not cover the whole body at a time, the bed must move to ensure the full coverage of the volume.

As was mentioned before, faster acquisitions result in noisier images. The standard scan-time per AFOV considered in this work is 70 seconds, as per the prevailing clinical protocol. Fig. 2.2 displays three simulated fast PET scans, with 15, 20, 30 seconds per AFOV, and the considered standard, 70 seconds per AFOV. Table A.3 of appendix A displays a close-up of a region in the coronal and sagittal planes of a given patient, for better visual assessment.



(a) 15 s/AFOV (b) 20 s/AFOV (c) 30 s/AFOV (d) 70 s/AFOV

Figure 2.2: Whole-body [^{18}F]FDG PET scans acquired with 15, 20, 30 and 70 s/AFOV, from left to right, respectively. PET/CT scans were performed on a Philips Vereos Digital PET/CT in the Champalimaud Clinical Centre.

It is possible to observe that, as expected, the shorter the time per AFOV, the worse the image quality. The idea is to find a denoising method able to restore the fast-scans to an image quality as close as possible to the standard-acquisition scan. To this end, the denoised fast scans will be compared, quantitatively, to the respective standard 70 s/AFOV scans, which will, therefore, serve as the reference.

3 | Denoising Methods

3.1 Standard Methods

Any type of image acquisition is susceptible to noise. Especially in a clinical environment, where there are greater limitations regarding not only technology but also patient-care concerns, images tend to display artefacts that may condemn the final result. As a consequence, the term *denoising* is not unknown to medical imaging. As was previously mentioned, one of the most common techniques used is Gaussian filtering, given its simplicity and efficiency in smoothing images [18].

3.1.1 Gaussian Filter

In image processing, smoothing a given image with a Gaussian filter is to perform a convolution between the image and a Gaussian function. Since an image consists of an array of values, to apply a convolution, this function must be translated into its discrete approximation. This is, the Gaussian function must be transposed into a kernel, which must be truncated, generally at three or more standard deviations from the mean of the chosen Gaussian distribution. The kernel will, then, slide over the image and the operation is carried out.

The filter's full width at half maximum (FWHM), in millimeters, will depend on the size of the image's pixels (or voxels), also in millimeters. This parameter establishes a threshold for the low-pass spatial filter, i.e., it defines the degree of smoothness of the filter. FWHM, by definition, depends on the standard deviation of the Gaussian distribution:

$$\text{FWHM}_{[\text{px}]} \cdot \text{pixel width} = \text{FWHM}_{[\text{mm}]} \Leftrightarrow 2\sqrt{2\ln 2}\sigma \cdot \text{pixel width} = \text{FWHM}_{[\text{mm}]} \tag{3.1}$$

3.2 Deep Learning

Deep learning is an AI technique that has demonstrated to be a promising tool in building denoising algorithms for clinical imaging, as the literature suggests [19]. DL is a branch of machine learning that uses artificial neural networks (ANNs) to learn intricate concepts from simpler ones [20]. ANNs consist of multiple layers of artificial neurons, able to extract complex features from raw input and learn them through adjusted weights and biases. These networks are an attempt to mimic the brain's neural networks into a computational model.

3.2.1 Activation Function

Mathematically, an artificial neuron introduces the concept of non-linearity to an otherwise linear model. A deep artificial neural network exhibits hidden layers, apart from the input and output layers. The characteristic of non-linearity takes the form of an activation function, in the hidden layers.

The rectified linear unit (ReLU) [21] is one of the most common activation functions, given its simplicity and efficient computation, while performing well on deep neural networks.

3.2.2 Loss Function

Broadly, neural networks can be trained in a supervised or unsupervised way. In medical imaging, the problem typically requires a discriminative (supervised) model, in which there are inputs and corresponding labels (targets). The process of learning consists of training a neural network that outputs a label-like object, given its input. To this end, a loss function is necessary to measure how far the prediction of the model is from the real label. The goal of learning is, then, to find the network parameters that minimise this loss function [22]. To achieve this, a so-called optimiser (described below, subsection 3.2.3) is used.

For image-like inputs, a relevant loss function is the mean squared error (MSE) between image and target. When employing this function, the optimiser's goal is to minimise the MSE between the training image set and respective reference images.

3.2.3 Optimiser

When training an ANN, an optimisation strategy must be implemented to determine which network parameters minimise the losses. This strategy usually starts by setting the network's parameters to random values (random initialisation), and then iteratively improving them, with each iteration being an attempt to decrease the given loss function [20]. The size of the steps is determined by the learning rate. If this hyperparameter is too small, then the algorithm must perform too many iterations to reach the desired minimum and training will be too long. If, however, the learning rate is too high, the algorithm may never converge as it jumps too far away at each iteration. It is, thereby, essential to find a compromise between these two extremes that allows the optimiser to efficiently converge to a minimum.

3.2.4 Convolutional Neural Networks

Convolutional neural networks (CNNs), or ConvNets, are derived from the study of the brain's visual cortex, being an attempt to replicate biological visual processing. CNNs are a type of ANNs and are most frequently used in image analysis, due to the fact that these networks expect an image-like input [22]. This accomplishes, in due course, a better encoding of certain image properties. Various architectures of CNNs have been developed and are being tested with different aims, particularly for image classification, object localisation and image segmentation.

The main aspect of ConvNets is the introduction of convolutional layers. Each of these involves the convolution of a kernel (filter) to the image matrix and is generally followed by a pooling layer to downsample it, as illustrated in fig. 3.1. This process is, essentially, feature extraction followed by dimensionality reduction.

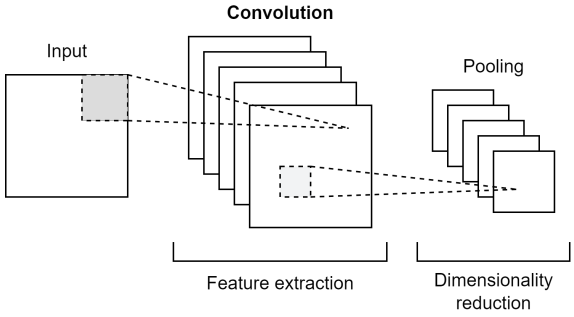


Figure 3.1: Illustration of a convolutional layer followed by a pooling layer.

Some convolutional networks are employed for simple classification tasks, where the output, given the image-like input, is the corresponding predicted label. However, in medical imaging processing tasks, the classification process should include localisation, i.e., a class should be assigned to each pixel. U-Net is a U-shaped CNN architecture, built specifically for biomedical image segmentation [23]. The U-shape is a result of the encoder-decoder property of the U-Net, in which the input image is downsampled in the so-called contracting path and upsampled in the expansive path. This way, the output is also an image of the same size as the input image. Symmetrical connections between downsampling and upsampling are established to perform the upsampling task. A schematic U-Net is represented in fig. 3.2.

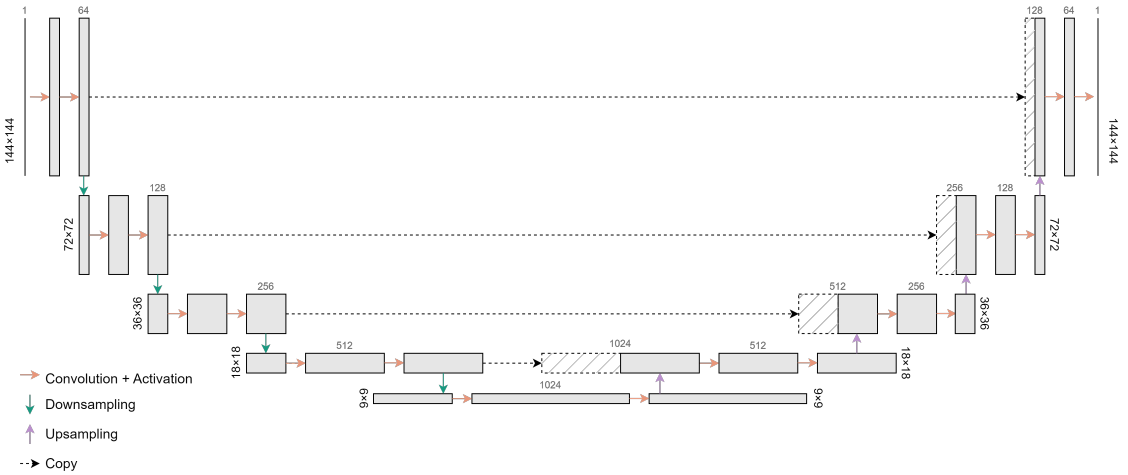


Figure 3.2: Schematic of a two-dimensional (2D) U-Net. The input is a 144×144 px image. Each grey square corresponds to a multi-channel feature map. Network architecture is based on the one described by Ronneberger, Fischer and Brox [23].

4 | Materials and Methods

The methodology used to undertake this work can be divided into three groups. The first one is related to acquiring the images used to train and test the network. The second refers to the implementation of the proposed DL-denoising method and the auxiliary one used for benchmarking. The latter constitutes the quantitative analysis employed to evaluate performance and to establish a comparison between different networks, as well as between the chosen network and standard denoising.

4.1 Dataset

The scans included in the dataset were conducted on patients with various types of cancer, referred for clinical [^{18}F]FDG PET/CT examinations in the Champalimaud Clinical Centre. The comprised whole-body [^{18}F]FDG PET/CT acquisitions were performed in a Philips Vereos PET/CT scanner, approximately 60 minutes post-injection, and with an acquisition time of 70 seconds per AFOV. To simulate fast scans, the already acquired data was reconstructed again, with only a portion of the total counts, extracted from the middle of the acquisition time-frame. The protocol for all reconstructions fulfills EANM Research Ltd. (EARL) ^{18}F standards 1 accreditation [16] (table 4.1).

Table 4.1: Reconstruction parameters of the Philips Vereos Digital PET/CT.

Parameter	Philips Vereos Digital PET/CT
Voxel volume	$4 \times 4 \times 4$ mm^3
Reconstruction algorithm	OSEM
Iterations	3
Subsets	15
Relaxation factor	1.0
Post-reconstruction filter	None, 3 mm or 5 mm

A coronal plane from each of the simulated fast scans and from the standard exam of a patient are shown in fig. 2.2.

The number of counts detected by the PET scanner is, in the described conditions, approximately proportional to the acquisition time. This means that reducing the number of counts to, for instance, half, would result in an image (nearly) identical to the one that would be obtained if the time-per-AFOV was halved.

112 whole-body [^{18}F]FDG PET scans from different patients were included in the dataset. 20 scans were randomly set aside for testing the deep learning models and the remaining 92 were employed in the training. The respective demographic characteristics are presented in table 4.2.

Table 4.2: Demographic characteristics of the patients included in the dataset.

Parameter	Total (112)	Training set (92)	Test set (20)
Gender (M/F)	48%/52%	50%/50%	40%/60%
Age [years]*	65 ± 11	65 ± 11	64 ± 10
Height [m]*	1.66 ± 0.09	1.66 ± 0.09	1.63 ± 0.09
Weight [kg]*	72 ± 15	71 ± 13	76 ± 19
BMI [kg/m^2]*	26 ± 5	26 ± 4	28 ± 5

* Displayed as average value ± standard deviation.

The average injected activity per patient included in the dataset was 244 MBq, with a standard deviation of 50 MBq. This corresponds to an average of 3.4 ± 0.2 MBq/kg. From within the training and test sets, the average injected activity was 3.4 ± 0.2 MBq/kg and 3.2 ± 0.3 MBq/kg, respectively.

Fig. 4.1 displays the occurrence of the different primary tumours in the dataset.

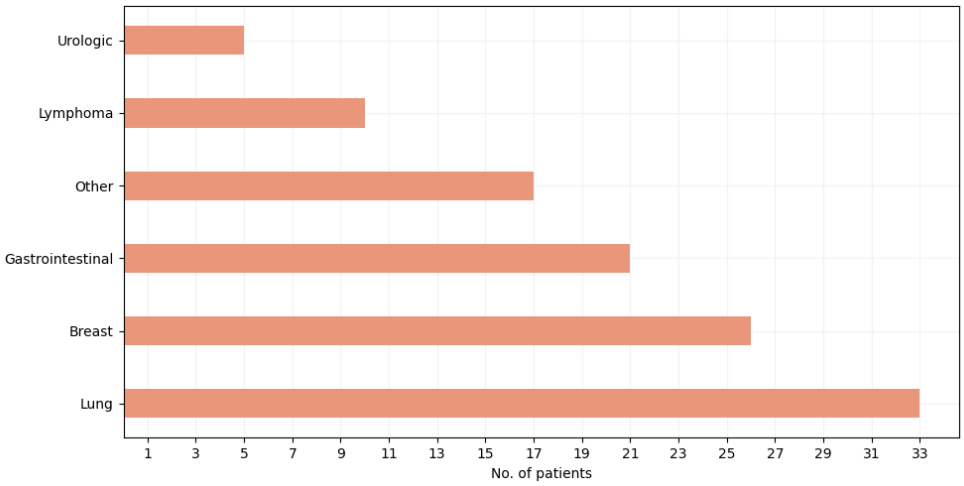


Figure 4.1: Bar plot of the primary tumour type in the dataset (112 patients) and the number of occurrences.

From the 20 PET/CT scans used to test the network, a total of 76 lesions were identified by clinicians in the reference image set (70 s/AFOV). The distribution of the metabolic tumour

volume of the identified lesions is shown in fig. 4.2.

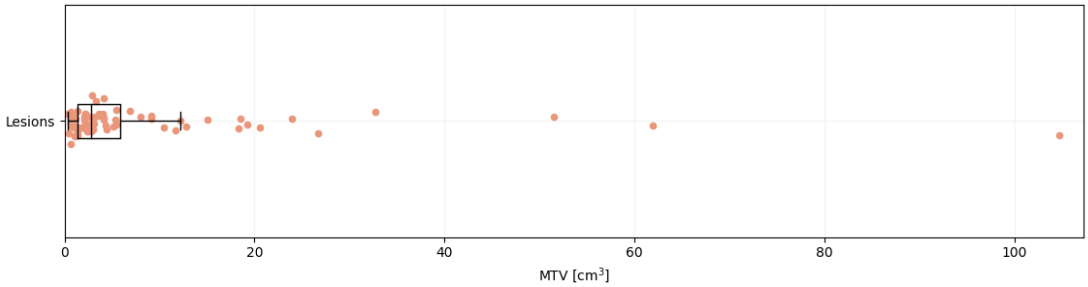


Figure 4.2: Box plot of the metabolic tumour volume (MTV) of the 76 lesions identified in the (reference) test set.

4.2 Deep-learning-based Approach

The approach to address fast PET denoising was based on deep learning. This requires specific software and hardware. For the implementation of the CNN, the machine learning Python framework *PyTorch* provided the building blocks. For training, validating and testing purposes, a specific graphics processing unit (GPU) allowed the use of the NVIDIA *CuDNN* library, a GPU-accelerated library for deep neural networks.

4.2.1 2D U-Net

In the initial stage, a 2D U-Net was implemented, for the denoising of the 3D PET images by denoising individually each axial slice. A simplified schematic of the network's architecture is illustrated in fig. 3.2.

In the encoder path, each block consists of two 3×3 convolutions (each followed by a ReLU activation), followed, in turn, by an average pooling that halves the input's size (down-sampling). In the input block, the first convolution outputs 64 feature maps from the input image. Excluding the input block, the first convolution of each block doubles the number of feature maps (channels), until the decoder path is reached.

In the decoder path, each block consists of a 2×2 transposed convolution for upsampling the block's input image, followed by a concatenation with the result of the symmetric block in the encoder layer, and a convolutional block of two 3×3 convolutions (each followed by a ReLU activation).

Weight initialisation for the convolutions in the convolutional blocks was He initialisation [24]. For the transposed convolution in the upsampling process, all weights were initialised with zero, as well as the biases in every instance.

The architecture of this network was designed keeping in mind its aim. As denoising is not a classification task, average pooling was chosen over max pooling. Batch normalisation, usually employed in segmentation tasks, was not performed as the scale of intensities of the PET images is to be maintained.

4.2.2 2.5D U-Net

Having in view the inclusion of spatial context information, a 2.5D U-Net was implemented, for the denoising of the 3D PET images by denoising individually each axial, coronal and sagittal slice. To this end, for training, each volume was split into all its planes, in each of the three main directions. The network's architecture is the same as described in subsection 4.2.1. As the network's input image size was 144×144 pixels, the axial slices were included fully-sized, while the coronal and sagittal slices were split into patches of this size.

4.2.3 3D U-Net

A 3D U-Net was also implemented taking into consideration the importance of spatial context information in the training. Contrary to the previous approaches, the 3D network's input is a 3D volume. Patches of $128 \times 128 \times 128$ voxels were fed into the network. Concerning network architecture, although the composition of the convolution blocks was kept, the number of blocks of the U-Net was reduced to three to decrease computational complexity. The resulting network architecture is illustrated in fig. 4.3.

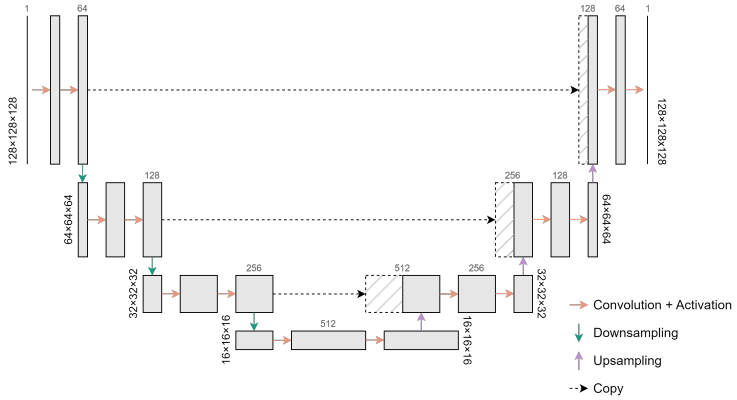


Figure 4.3: Schematic of a 3D U-Net. The input is a 128×128×128 vx patch extracted from the original image. Each grey square corresponds to a multi-channel feature map.

Weight initialisation was equivalent to the strategy employed when implementing the 2D and 2.5D U-Nets: He initialisation [24] for the convolutions in the convolutional blocks, and zero initialisation for the transposed convolution in the upsampling process, as well as for the biases in every instance.

4.2.4 Training and Testing

From the 112 whole-body [^{18}F]FDG PET scans, 82 were used for training, 10 were used for validation during training and 20 were used for testing. The loss function employed in the training of the networks was the mean squared error. Adam [25] was the chosen optimiser. The starting learning rate was 0.0005, and every 50 epochs it was lowered by 10%.

When training the 2D U-Net, every axial plane from each 3D volume was fed into the network per epoch. For the 2.5D network, a random selection of 10% of the planes in each of

the three main directions (axial, coronal and sagittal) was the network's input in each epoch. For the training of the 3D U-net, one single patch randomly extracted from each image was fed into the network per epoch. The default number of epochs to train the models was 1000. Every time the loss reached a minimum within the internal validation set in a given epoch, the associated network was saved.

4.2.4.1 Data Augmentation

Data augmentation is a technique often used to reduce overfitting when training a model. It consists of generating modified data from the original, increasing artificially the training set. Data augmentation was performed through Python's open-source library *SimpleITK*.

In the training of the 2D and the 2.5D U-Nets, the same data augmentation methods were employed, as the networks' inputs were similar (144×144 px images). These consisted of rotation, scaling, translation and noise addition. The rotation angle was obtained by extracting a random value from a normal distribution with 0.0° mean and 2.5° standard deviation. Parameters for anisotropic scaling were randomly extracted from a normal distribution of mean 1.00 and standard deviation 0.02. For the translation, an offset was randomly extracted from a normal distribution of mean 0 mm and standard deviation 50 mm for the two directions. Multiplicative noise in the form of a matrix of values extracted from a normal distribution of mean 1.0 and standard deviation 0.1 was applied to the input images.

In the training of the 3D U-Net, rotation, scaling and noise addition were performed. The rotation angle in each direction and the parameters for anisotropic scaling were obtained as before. No translation was performed, as the 3D patches used in the training were already randomly selected from the original full volume. Random noise was also applied to each input training image, as before.

4.3 Gaussian Filter

Python's open-source library *scipy* provided a central processing unit (CPU)-based implementation of the Gaussian filter. As was introduced in subsection 3.1.1, the filter's FWHM depends on the given Gaussian distribution's standard deviation. Accordingly, and taking into account eq. 3.1, for a given pixel/voxel width and by providing the desired standard deviation, a specific-width filter is defined. The Gaussian kernel was truncated at three standard deviations from the mean of the chosen Gaussian distribution, e.g. a $\text{FWHM}_{[\text{mm}]}$ of 5 mm corresponds to a standard deviation of 0.53×4 mm (given eq. 3.1 and for a voxel width of 4 mm, as specified in table 4.1) which, in turn, results in a kernel of size $2 \times (0.53 \times 3) + 1 = 5$ vx.

4.3.1 Optimisation

Considering a $4 \times 4 \times 4$ mm voxel (as specified in table 4.1), different filter widths were tested having in view the minimisation of the MSE between the images obtained from the fast and the reference scans. The filter widths tested for each scan duration ranged from 1 to 12 mm.

The different filters were obtained by applying successive increments of 0.5 mm to the starting 1 mm filter width.

4.4 Performance Evaluation

The performance of the denoising methods was determined by how close their output image sets were to the reference image set. Several measures were used to perform this quantitative evaluation: voxel-wise and region of interest (ROI) quantification. Both are described in the following subsections.

4.4.1 Voxel-wise Analysis

The most straightforward comparison between two images is by observing them as a whole, i.e., taking into account the values of every corresponding pixel/voxel in the images. This is what is meant by a voxel-wise analysis.

When interpreting this analysis, performed globally in the PET images, it's important to take into consideration the great number of background voxels in the volume. These will carry a value equal to or very close to zero and are expected to match between reference, noisy and denoised images. As a result, the voxel-wise measures are expected to reveal a strong agreement, from the outset, between the fast and reference images.

4.4.1.1 Mean Squared Error

As the name suggests, mean squared error (MSE) measures the average of the squares of the errors (differences) between two sets of values. MSE was implemented using Python's open-source image processing library *scikit-image*.

4.4.1.2 Structural Similarity Index Measure

The structural similarity index measure (SSIM) reveals how similar two images are. It is typically employed to assess the quality of a given image, when compared to its full-quality equivalent. The SSIM is usually computed on various windows of an image, considering three comparison measurements between the samples: luminance (or intensity), contrast and structure [26]. The resulting formula reflects these three properties in a weighted combination. In this study's context, SSIM was obtained for the whole image.

SSIM is a value comprised between 0 and 1, being a SSIM equal to 1 an indication of perfect similarity between the images.

It was implemented using Python's open-source image processing library *scikit-image*.

4.4.1.3 Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC) is a measure of reliability between sets of values, i.e., a measure of how consistent two sets of values are. To quantitatively compare two images

through the intraclass correlation coefficient (ICC), a voxel-wise analysis is performed: the voxel-values from one image against the corresponding voxel-values in the other image.

Three different types of ICC were described by Shrout and Fleiss [27]. In the reliability study in question here, the PET image is the so-called target and the methods being compared (each of the denoising methods implemented and the standard-quality acquisition) are the so-called judges. The different outputs of these two methods are the ratings, which will be the objects for the ICC comparison. The closer the ICC is to 1, the better the quantitative agreement between the images.

The ICC for absolute agreement, $ICC_{(2,\,1)}$ as described in [27], implemented in Python, was utilised in the voxel-wise analysis.

4.4.1.4 Statistical Analysis

The non-parametric Friedman test for repeated measures was used to assess statistically significant differences among denoising techniques. The null hypothesis states that all the samples being compared have the same distribution. If the null hypothesis is rejected, i.e., if the *p*-value is significant, it can be assumed that at least two samples come from different distributions. Post hoc analysis through the Wilcoxon signed-rank test was used, subsequently, to discriminate specifically the differences between pairs of samples. The Bonferroni correction was performed when running the post hoc tests.

For all statistical tests performed, a significance level of 5% was established.

Python's open-source library *scipy* provides implementations of both the Friedman and the Wilcoxon signed-rank tests, which were employed for the respective tasks.

4.4.2 Regional Quantification Analysis

Besides the voxel-wise analysis, a ROI-based analysis was performed. The regions considered to this end were the normal-uptake organs and the malignant lesions identified by the clinicians in the reference images.

4.4.2.1 Normal-Uptake Organ Quantification Analysis

A powerful indicator of image quality is the signal-to-noise ratio (SNR). Broadly, SNR is a measure of how prominent a given signal is compared to the existing background noise. For imaging purposes, SNR is routinely achieved by the quotient of the average signal value and the corresponding standard deviation. In the present study, this quotient was obtained not for the image as a whole, but for the regions in which uptake uniformity is expected. Thus, the delineation of such was carried out in a subset of the PET scans included in this study. This task was done manually in the reference images, by selecting at least 100 voxels located approximately in the same region (of the liver and lungs) across patients. The resulting mask for a given patient was the same for quantification in each set of images (original, denoised and reference). This made it possible to compute the SNR in the liver and lungs, through the quotient of the mean SUV in the given region and the corresponding standard deviation.

4.4.2.2 Tumour Quantification Analysis

As was mentioned above, another key consideration in the quantitative analysis is regarding lesion identification and quantification in the denoised fast scans. For these images to be considered clinically viable, the same lesions identified in the reference images must be identified in the denoised images. Thus, for each set of images, semi-automatic Bayesian segmentation [28], [29] was performed from masks (marked ROIs) outlined by the clinicians in the reference images. Following this segmentation, feature extraction took place and is described in greater detail below. A given lesion was only considered in the quantification analysis if its segmentation resulted in a volume equal or superior to 4 voxels (0.256 cm^3, given the specifications of the PET/CT scanner in table 4.1).

To verify if the lesion features are maintained (or not) after subjecting the image through a given post-processing method, different measures are commonly used. One is to observe if the maximum SUV (SUV_{max}) of a given ROI is similar in both the resulting and the original images. Another is to compute the mean standardised uptake value (SUV_{mean}) in a lesion and, once again, check if it's similar between both images. These measures and four additional ones were also utilised for lesion quantification — standardised uptake value standard deviation (SUV_{SD}), peak standardised uptake value (SUV_{peak}), total lesion glycolysis (TLG) and metabolic tumour volume (MTV) — are described in table 4.3.

The ideal agreement between reference and denoised images would be if the SUVs matched between images. Hence, the closer the observed SUV parameter of the denoised image is to the corresponding standard image, the better the denoising method.

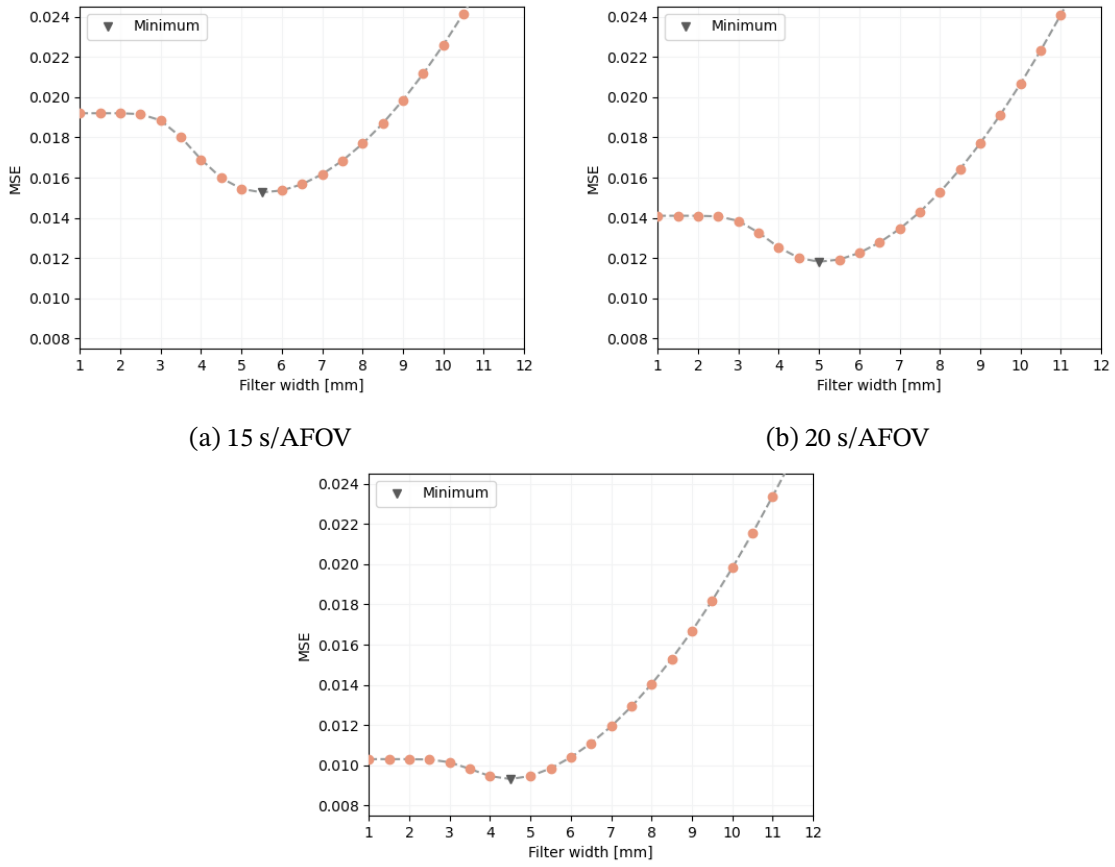
Table 4.3: Measures used for lesion quantification.

Measure	Method
Maximum standardised uptake value (SUV_{max})	Maximum value of the set of the lesion's voxels' intensities.
Mean standardised uptake value (SUV_{mean})	Mean value of the set of the lesion's voxels' intensities.
Standardised uptake value standard deviation (SUV_{SD})	Measure of dispersion of the lesion's voxels' intensities, relatively to the average value.
Peak standardised uptake value (SUV_{peak})	Maximum average intensity of a 3×3×3 voxel volume within the lesion.
Total lesion glycolysis (TLG)	Product of the voxel's volume and the sum of the set of the lesion's voxels' intensities.
Metabolic tumour volume (MTV)	Product of the number of voxels in the lesion and the voxel's volume.

5 | Results

5.1 Gaussian Filter

As described in section 4.3, the Gaussian filter was optimised in terms of width, through the minimisation of MSE relatively to the reference images. For the Gaussian filter (GF)-denoised 15, 20 and 30 s/AFOV images, the optimal filter widths achieved, for the training set, were of 5.5, 5 and 4.5 mm, respectively, as is shown in fig. 5.1.



(a) 15 s/AFOV

(b) 20 s/AFOV

(c) 30 s/AFOV

Figure 5.1: Gaussian filter's width optimisation results. Average MSE between the GF-denoised images and the corresponding reference images, from within the training set, per filter width, for each scan duration.

For each scan duration, the respective optimal filter was applied to the original (not denoised) test set, and the resulting images were used as benchmarks in the following sections.

Table 5.1 exhibits the voxel-wise analysis of these GF-denoised sets. A slight but significant improvement (p<0.001), when compared with the original fast scans, was observed regarding MSE, ICC and SSIM, for the three scan durations.

An improvement was also observed regarding SNR and SUV_{mean} compared to the reference in the liver and lungs, when compared with the original fast images, as per table 5.2. This is, the GF-denoised sets revealed smaller differences relatively to the reference than the original scans.

Lesion quantification is described by the the 95% limits of agreement with the reference, the median absolute deviations from the reference and the relative differences to the reference (tables 5.3, 5.4 and A.2 (appendix A), respectively), regarding SUV_{max}, SUV_{mean}, SUV_{SD}, SUV_{peak}, TLG and MTV.

5.2 2D U-Net

The training progress of the 2D U-Net for each scan duration is illustrated through the plots in fig. 5.2. The networks that output the validation loss minimums, for the 15, 20 and 30 s/AFOV training sets, corresponded to the 203rd, 303rd and 236th epochs, respectively. Since the training entered an overfitting pattern in which the validation loss worsened, for representation purposes, in fig. 5.2, the last epoch shown is the 750th.

Table 5.5 displays a coronal plane from an image of the test set, denoised through the 2D U-Net, for each scan duration, along with the remaining denoising results.

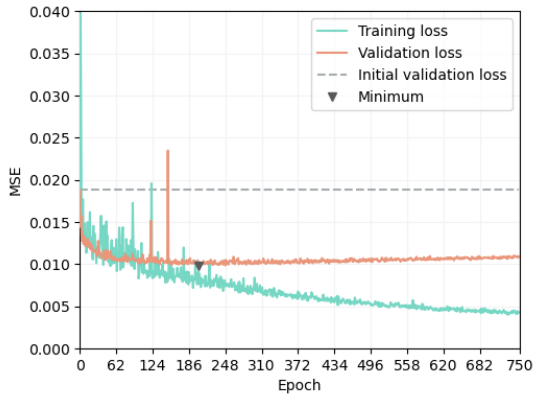
5.2.1 Voxel-wise Analysis

The voxel-wise analysis was performed between the reference images and each of the original and denoised test sets. Results for the 2D U-Net denoising are presented in table 5.1. Statistically significant differences in terms of MSE, SSIM and ICC was observed between each pair of image sets (p<0.001). These differences reveal an improvement in the voxel-wise measures for each denoised set compared to the original, and for DL-denoising compared to GF-denoising.

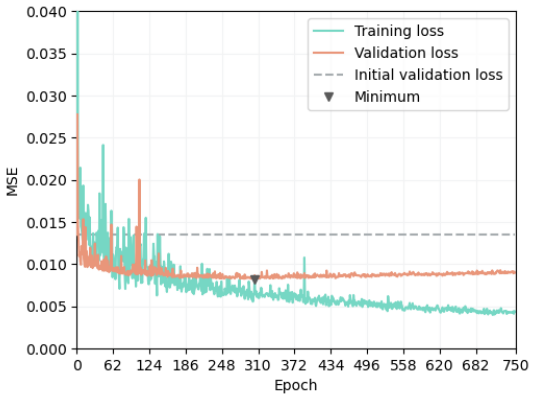
Box and parallel coordinates plots regarding MSE, SSIM and ICC in the test set are displayed in figs. A.1, A.2 and A.3 of appendix A, comparing the original fast scans and the two denoising methods, for the 15, 20 and 30-s/AFOV-based sets, respectively.

5.2.2 Regional Quantification Analysis

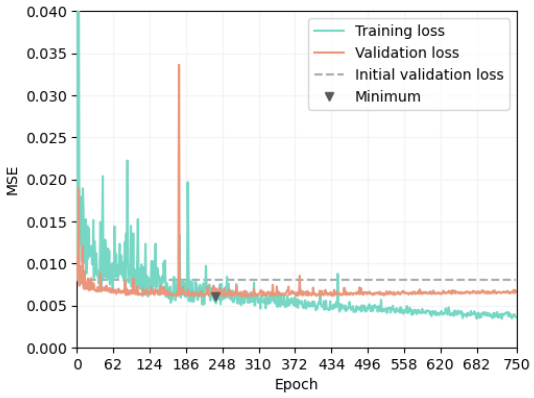
The quantification results for the 2D U-Net denoising are summarised in table 5.2 for the liver and lungs, and table 5.3 for the included lesions. An increase in SNR in both the liver and lungs was observed, compared to the reference images, while a decrease is observed in the original (not denoised) and Gaussian-filter-denoised sets. Regarding SUV_{mean} in the same regions, no remarkable difference was observed.



(a) 15 s/AFOV



(b) 20 s/AFOV



(c) 30 s/AFOV

Figure 5.2: Training progress of the 2D U-Nets for each scan duration. Plot of the training and validation losses (MSE) per epoch. Epoch 0 corresponds to the initial loss for the training and validation sets, before applying the network.

As to the lesions, the SUV_{max} 95% limits of agreement between the reference and 2D U-Net denoising reveal a dispersion of about 2 g/mL decrease or 1 g/mL increase, which corresponds to an interval similar to that for the original and the GF-denoised sets. Regarding MTV, the 95% limits of agreement with the reference showed a decrease or an increase no larger than 3 cm^3, which is not too different from the values observed for the original and the GF-denoised sets. The most pronounced discrepancies are reported for the 15-s/AFOV-based set.

The median absolute deviations from the lesions' features are presented in table 5.4. Table A.2 (appendix A) displays the relative differences to the reference images regarding lesion features, for the denoising through the 2D U-Net.

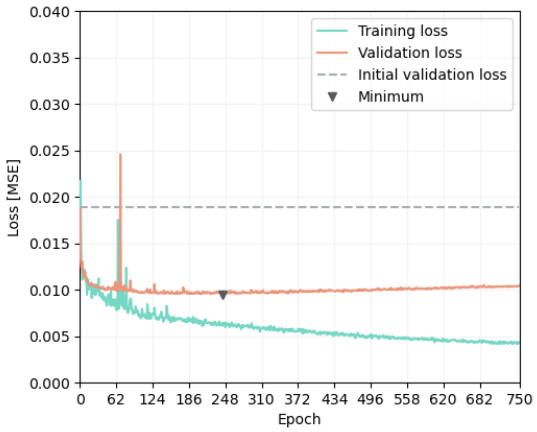
Although the quantitative analysis of the denoising through the 2D U-Net exhibits promising results, the output images reveal artefacts in the coronal and sagittal planes, arising from the lack of 3D spatial information when applying the network only to the axial planes, as can be seen in table A.1 (appendix A). The coronal and sagittal planes display what can be described as an "horizontal blurring", given that each horizontal line in these views corresponds to an

axial plane in 3D space.

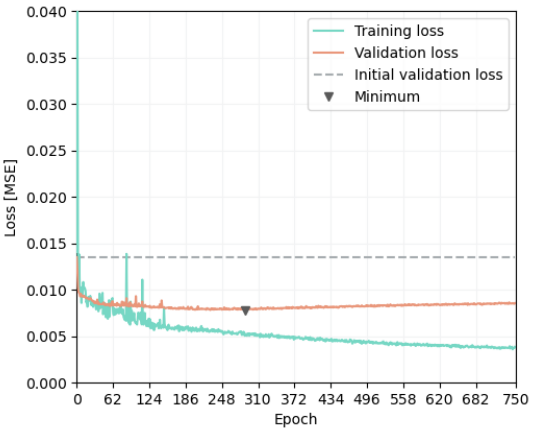
5.3 2.5D U-Net

The training loss of the 2.5D U-Net for each scan duration is recounted by the plots of fig. 5.3. The best models, associated with the absolute validation loss minimums, were saved on the 243rd, 288th and 354th epochs, for the 15, 20 and 30-s/AFOV-based networks, respectively. Once again, the training reached overfitting with worsening of the validation loss. Thus, for representation purposes, the last epoch shown in fig. 5.3 is the 750th epoch.

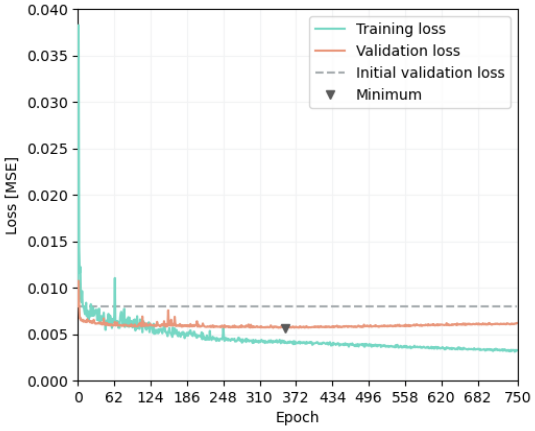
The output of the denoising of an image of the test set through the 2.5D U-Net for each scan duration in displayed in table 5.5, along with the remaining denoising results. A close-up of the abdominal region of a patient from the 15 s/AFOV test set is displayed in table A.3 (appendix A).



(a) 15 s/AFOV



(b) 20 s/AFOV



(c) 30 s/AFOV

Figure 5.3: Training progress of the 2.5D U-Nets for each scan duration. Plot of the training and validation losses (MSE) per epoch. Epoch 0 corresponds to the initial loss for the training and validation sets, before applying the network.

5.3.1 Voxel-wise Analysis

Statistically significant differences were observed in MSE, ICC and SSIM between the original, the GF-denoised and the DL-denoised (2.5D U-Net) images (p<0.001). Between each pair of image sets (original *versus* each denoising method and both denoising methods, one *versus* the other) significant improvement in favour of DL denoising was observed in terms of the three measures (p<0.001).

Figs. 5.4, 5.5 and 5.6 display the voxel-wise analysis results regarding MSE, SSIM and ICC between the different sets of images. Table 5.1 contains the associated average values. A decrease in MSE was observed for GF denoising compared to the original sets and, in turn, for DL denoising compared to GF denoising. Conversely, an increase in both SSIM and ICC was observed for GF denoising compared to the original sets and, in turn, for DL denoising compared to GF denoising.

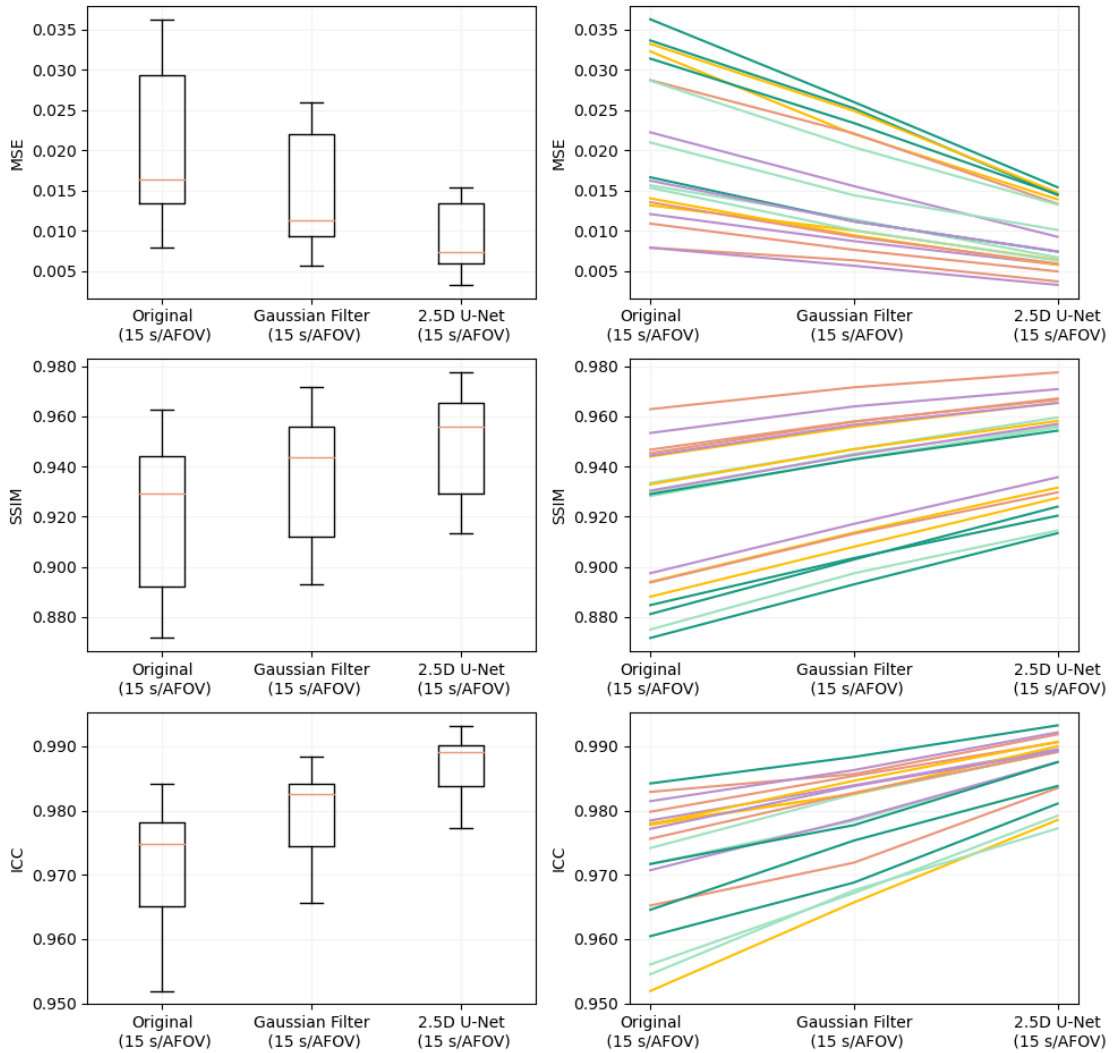


Figure 5.4: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 15 s/AFOV images, the GF-denoised images and the DL-denoised images (2.5D U-Net).

Each line in the parallel coordinates plots represents each test set image.

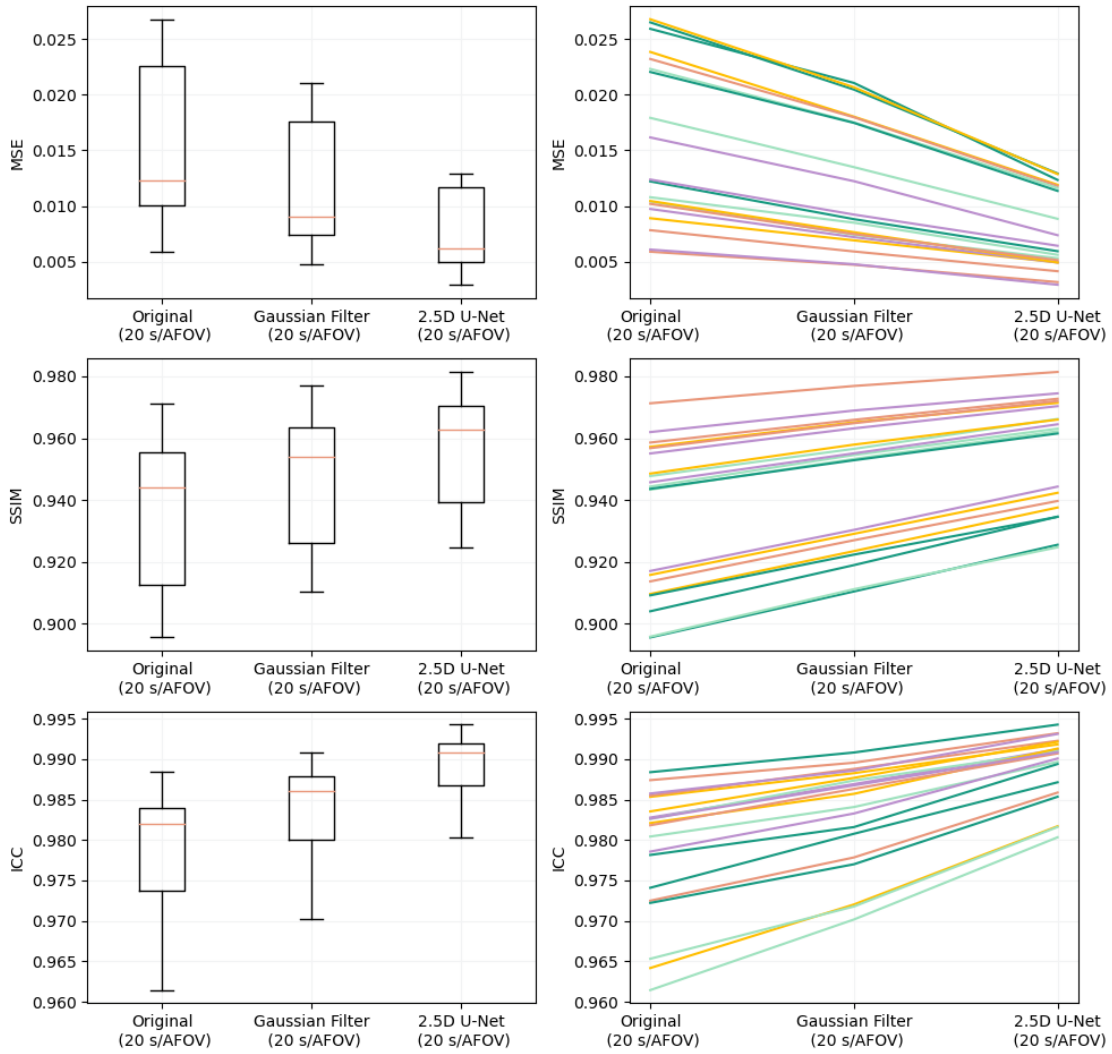


Figure 5.5: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 20 s/AFOV images, the GF-denoised images and the DL-denoised images (2.5D U-Net).

5.3.2 Regional Quantification Analysis

Table 5.2 showcases the 2.5D U-Net results of SNR and SUV_{mean} in the liver and lungs, for the three scan durations. The original and GF-denoised sets exhibit a decrease in SNR for both the liver and lungs relatively to the reference set. Contrariwise, DL denoising resulted in a noteworthy increase in SNR in the lungs and liver. Regarding SUV_{mean} in the three sets of images, an average variation of no more than 2% relatively to the reference set was observed, for all three scan durations, and both in the liver and lungs.

The segmentation of one of the lesions in the 30-s/AFOV-based set denoised through the 2.5D U-Net did not meet the criteria to be regarded in the quantitative analysis (volume inferior

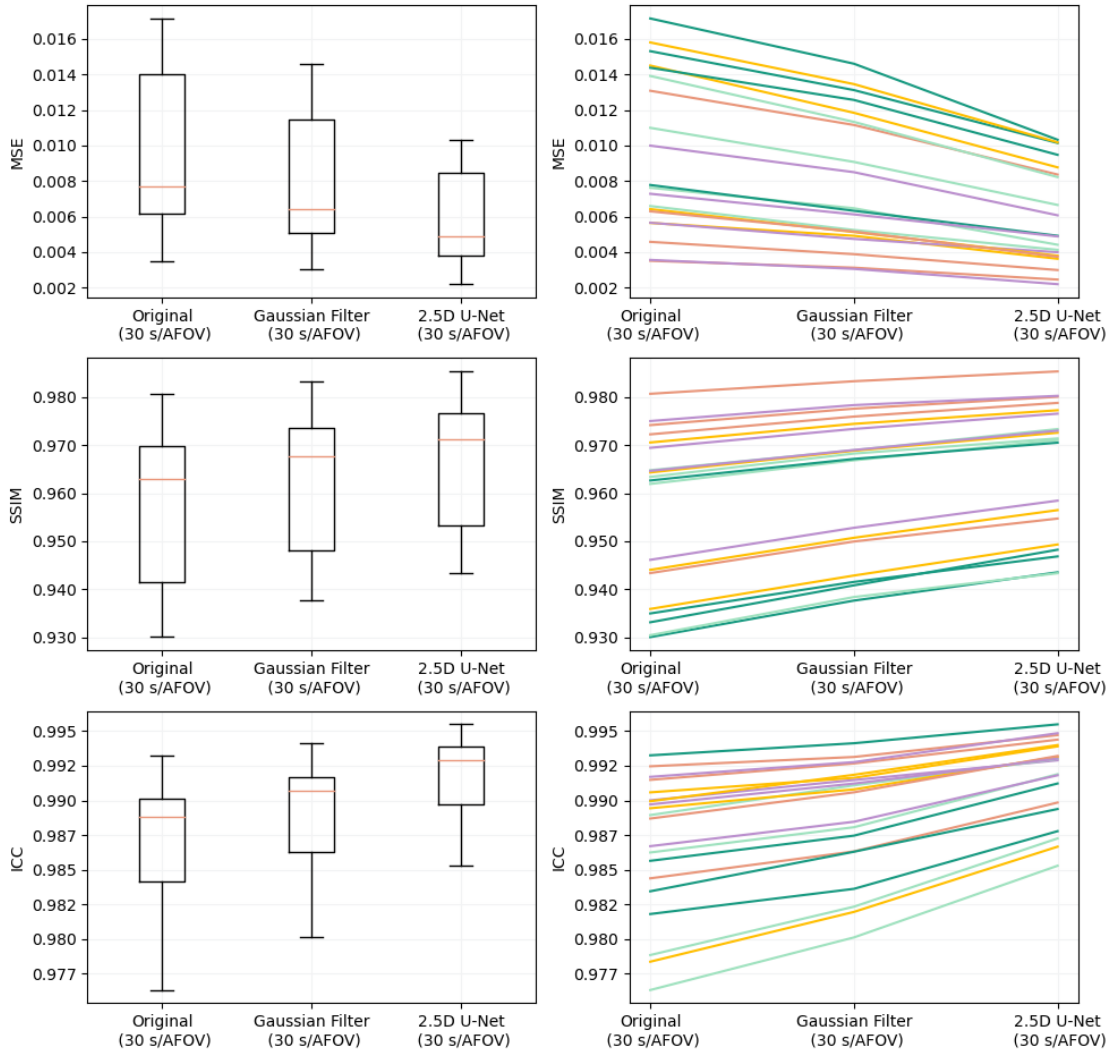


Figure 5.6: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 30 s/AFOV images, the GF-denoised images and the DL-denoised images (2.5D U-Net).

to 4 voxels, as per subsection 4.4.2.2). Thus, only 75 out of the 76 identified lesions were included in the quantification.

Table 5.3 displays the 2.5D U-Net results for the 95% limits of agreement concerning the extracted lesion features, for the three scan durations. In terms of SUV_{max}, a range of around 2 g/mL decrease or around 1 g/mL increase was recorded for all image sets, relatively to the reference full-duration scans. Regarding the remaining SUV measures, the 95% limits of agreement were similar between the original, GF and DL-denoised sets, with variations of about ± 1 g/mL for SUV_{mean} and SUV_{peak}, and even smaller for SUV_{SD}. No sharp variation was recorded in terms of MTV and, consequently, TLG, in the DL-denoised set. The 95% limits of agreement were approximately ± 3 cm^3 in the lesions' MTV. The broader dispersions in the lesions' features were recorded for the 15-s/AFOV-based sets.

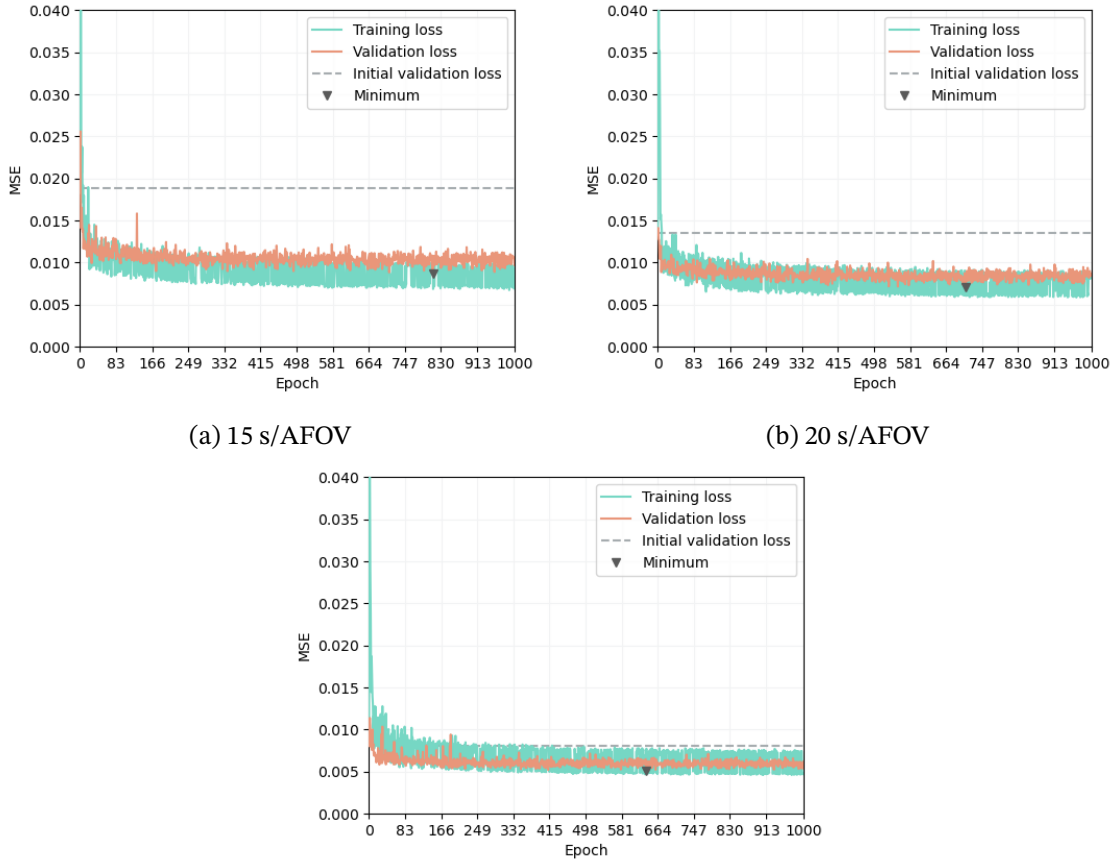
The median absolute deviations from the lesions' features in the reference images are presented in table 5.4. Table A.2 (appendix A) reports this lesion quantification analysis in terms of relative difference to the reference images, for DL denoising with the 2.5D U-Net.

Figs. A.5, A.6 and A.7 (appendix A) present the Bland-Altman plots for SUV_{max} and MTV, for the 15, 20 and 30-s/AFOV-based sets, respectively. These illustrate the results described above.

5.4 3D U-Net

Fig. 5.7 displays the training progress of the 3D U-Net for the three scan durations (15, 20 and 30 s/AFOV). As described in subsection 4.2.4, the network was saved every time a validation loss minimum was achieved. For the 15, 20 and 30-s/AFOV-based networks, the absolute minimum validation loss occurred for the 811th, 709th and 638th epochs, respectively.

Table 5.5 showcases the denoising through the 3D U-Net, for each scan duration, of an image of the test set, along with the remaining denoising results.



(a) 15 s/AFOV

(b) 20 s/AFOV

(c) 30 s/AFOV

Figure 5.7: Training progress of the 3D U-Net for each scan duration. Plot of the training and validation losses (MSE) per epoch. Epoch 0 corresponds to the initial loss for the training and validation sets, before applying the network.

5.4.1 Voxel-wise Analysis

Once again, there was statistically significant differences regarding MSE, SSIM and ICC between the original, GF-denoised and DL-denoised (3D U-Net) sets (p<0.001) and amongst each other (p<0.001), for the three scan durations considered.

Figs. 5.8, 5.9 and 5.10 show the 3D U-Net results of the voxel-wise analysis regarding MSE, SSIM and ICC from within the 15, 20 and 30-s/AFOV-based sets, respectively. The average values obtained in the test set for these three measures are displayed in table 5.1. A decrease in MSE and an increase in SSIM and ICC were observed, for the 3D U-Net compared to the original 15, 20 and 30 s/AFOV sets and the respective GF-denoised sets. To better discriminate the differences between the voxel-wise analysis of the denoising through the 2.5D and 3D U-Nets, the associated results are plotted side-by-side in fig. A.4 (appendix A), for the 20-s/AFOV-based image sets.

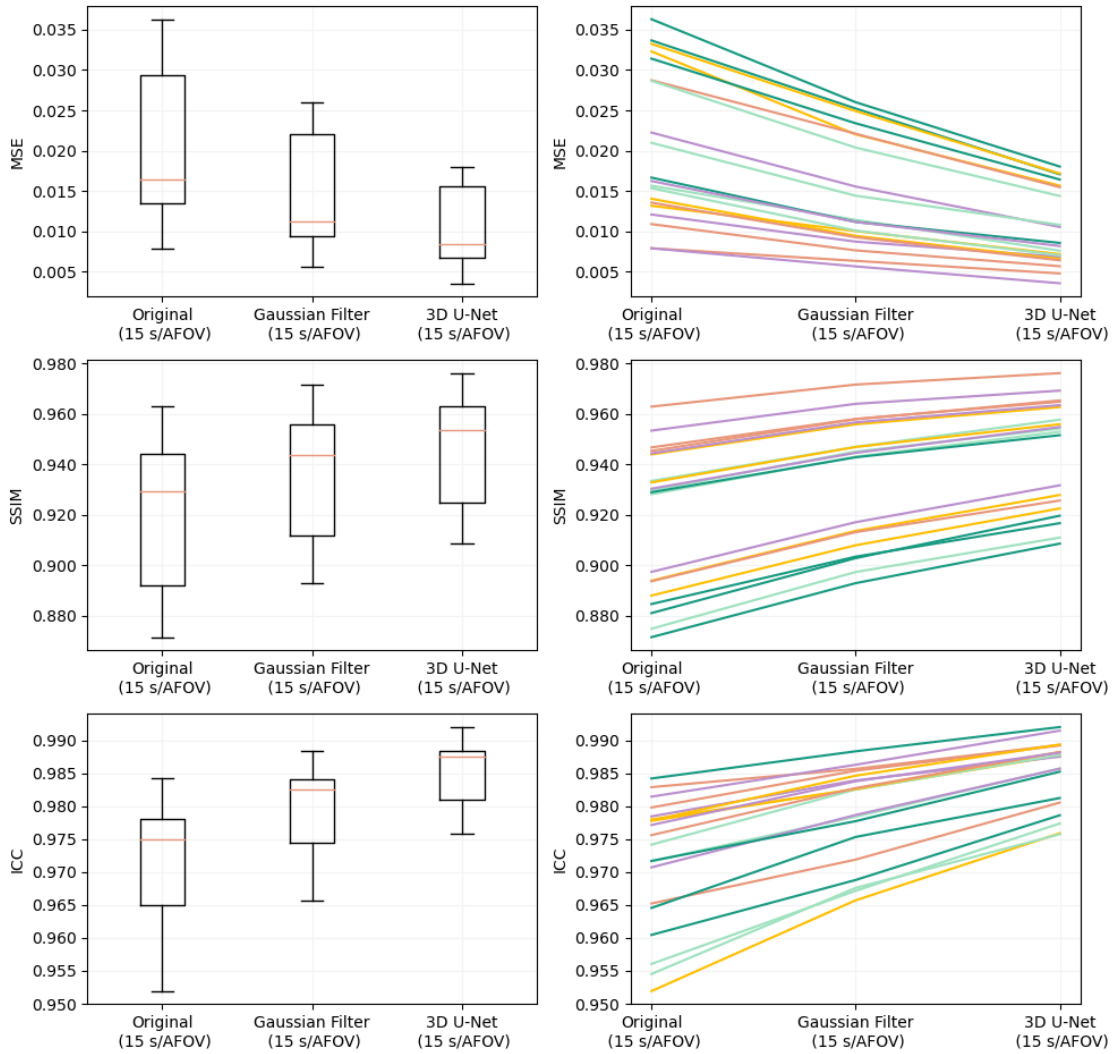


Figure 5.8: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 15 s/AFOV images, the GF-denoised images and the DL-denoised images (3D U-Net).

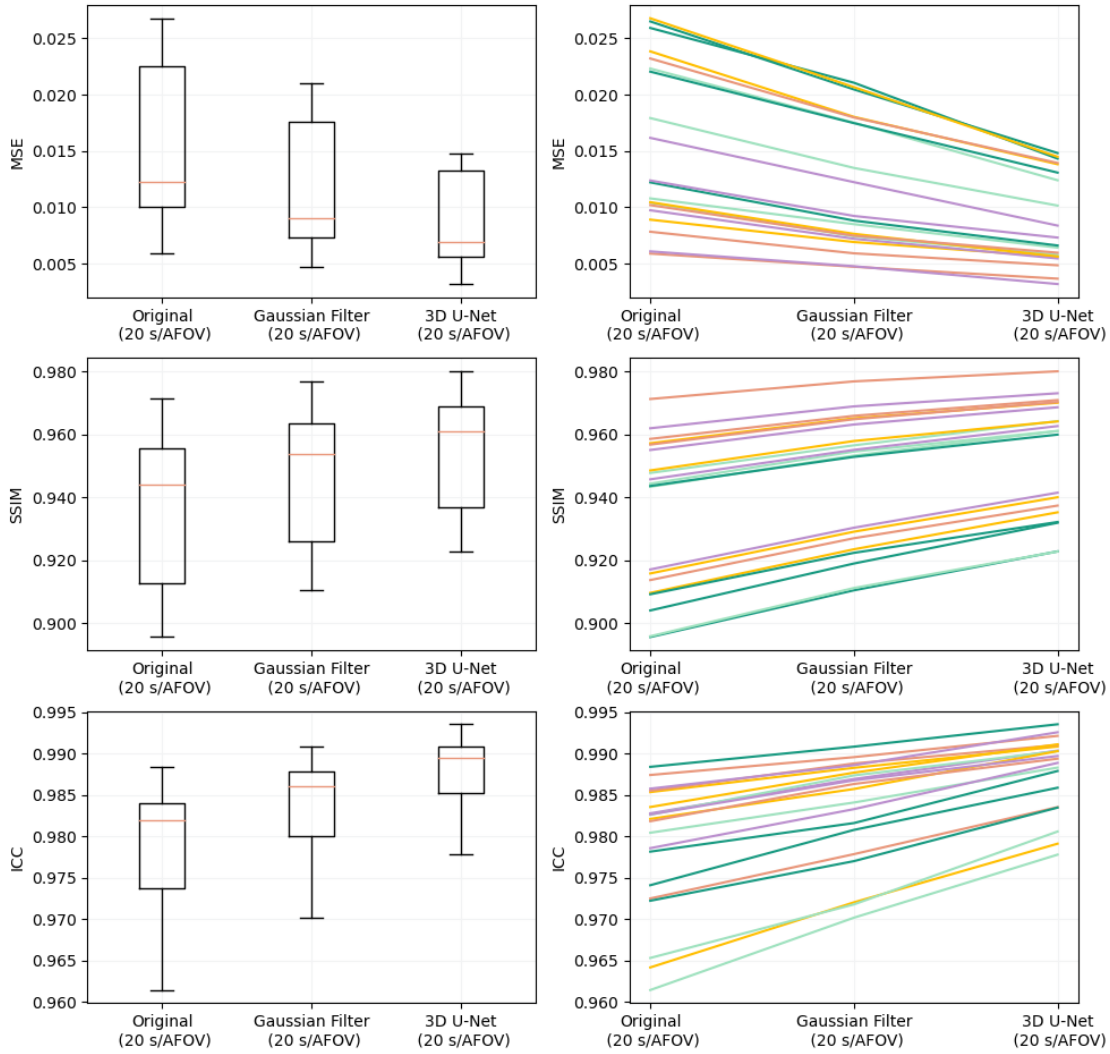


Figure 5.9: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 20 s/AFOV images, the GF-denoised images and the DL-denoised images (3D U-Net).

5.4.2 Regional Quantification Analysis

Table 5.2 sets forth the 3D U-Net results in terms of SNR and SUV_{mean} in the liver and lungs, for the three scan durations. Once again, an outstanding improvement was observed regarding SNR in both liver and lungs, for all scan durations considered. In terms of SUV_{mean}, there was a small difference compared to the reference images.

Tumour quantification results for the 75 included lesions are displayed in table 5.3. A similar variation in the lesions' SUV_{max} relatively to the reference images was observed for both GF denoising and DL denoising, as well as for the original sets, ranging between a decrease of 2 g/mL and an increase of 1 g/mL, for the three scan durations. The SUV_{peak} 95% limits of agreement for the three sets of images (not denoised, GF-denoised and denoised through the

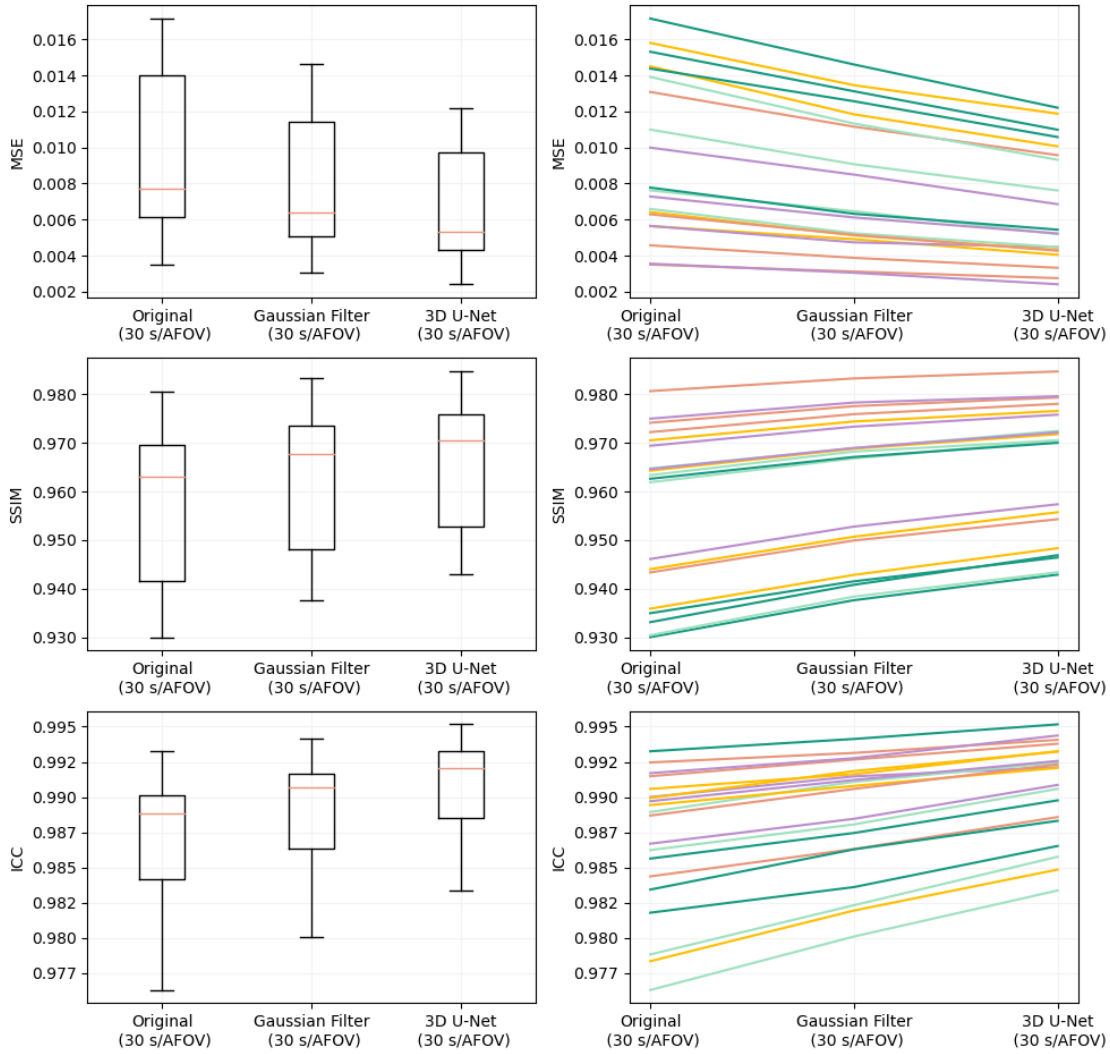


Figure 5.10: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 30 s/AFOV images, the GF-denoised images and the DL-denoised images (3D U-Net).

3D U-Net) revealed a variation of about $\pm$ 1 g/mL relatively to the reference. In terms of MTV, the limits were around ± 3 cm^3. The interval is narrower for the DL-denoised sets than for the original set in the 20-s/AFOV-based images, and for both the original and GF-denoised sets in the 15 and 30-s/AFOV-based images.

The median absolute deviations in the lesions' features are presented in table 5.4. The relative differences are presented in table A.2 (appendix A).

Bland-Altman plots with regards to SUV_{max} and MTV in the 75 included lesions for the 15, 20 and 30-s/AFOV-based sets denoised through the 3D U-Net are presented in figs. A.5, A.6 and A.7 of appendix A, respectively. These aid the interpretation of table 5.3.

Table 5.1: Average MSE, ICC and SSIM relative difference to the reference images (± standard deviation) for the original (not denoised) images and the images denoised with the different U-Nets and the benchmarking method (Gaussian filter).

Measure / Image set	MSE [$\times 10^{-3}$ g^2/mL^2]	SSIM	ICC
15 s/AFOV	21 ± 9	0.92 ± 0.03	0.972 ± 0.009
GF(15 s/AFOV)[(1)]	15 ± 7	0.93 ± 0.02	0.979 ± 0.007
2D U-Net(15 s/AFOV)[(2)]	11 ± 5	0.94 ± 0.02	0.985 ± 0.005
2.5D U-Net(15 s/AFOV)[(3)]	9 ± 4	0.95 ± 0.02	0.987 ± 0.005
3D U-Net(15 s/AFOV)[(4)]	10 ± 5	0.95 ± 0.02	0.985 ± 0.005
20 s/AFOV	15 ± 7	0.93 ± 0.02	0.979 ± 0.008
GF(20 s/AFOV)[(1)]	12 ± 6	0.95 ± 0.02	0.983 ± 0.006
2D U-Net(20 s/AFOV)[(2)]	9 ± 4	0.95 ± 0.02	0.988 ± 0.004
2.5D U-Net(20 s/AFOV)[(3)]	8 ± 3	0.96 ± 0.02	0.989 ± 0.004
3D U-Net(20 s/AFOV)[(4)]	9 ± 4	0.95 ± 0.02	0.988 ± 0.004
30 s/AFOV	9 ± 4	0.96 ± 0.02	0.987 ± 0.005
GF(30 s/AFOV)[(1)]	8 ± 4	0.96 ± 0.01	0.989 ± 0.004
2D U-Net(30 s/AFOV)[(2)]	7 ± 3	0.96 ± 0.01	0.990 ± 0.003
2.5D U-Net(30 s/AFOV)[(3)]	6 ± 3	0.97 ± 0.01	0.992 ± 0.003
3D U-Net(30 s/AFOV)[(4)]	7 ± 3	0.96 ± 0.01	0.991 ± 0.003

[(1)] Set denoised through Gaussian smoothing.
[(2)] Set denoised through the 2D U-Net.
[(3)] Set denoised through the 2.5D U-Net.
[(4)] Set denoised through the 3D U-Net.

Table 5.2: Average SNR and SUV_{mean} relative difference to the reference images (± standard deviation) for the liver and lungs in the original (not denoised) and denoised (Gaussian filter and U-Net) images.

ROI	Liver		Lungs	
Measure / Image set	ΔSNR [%]	ΔSUV_{mean} [%]	ΔSNR [%]	ΔSUV_{mean} [%]
15 s/AFOV	-49 ± 8	-2 ± 8	-41 ± 10	-5 ± 7
GF(15 s/AFOV)[1]	-27 ± 14	-2 ± 7	-20 ± 13	-5 ± 7
2D U-Net(15 s/AFOV)[2]	$+118 \pm 48$	0 ± 4	$+50 \pm 33$	$+2 \pm 6$
2.5D U-Net(15 s/AFOV)[3]	$+138 \pm 42$	$+1 \pm 5$	$+68 \pm 36$	$+4 \pm 7$
3D U-Net(15 s/AFOV)[4]	$+103 \pm 48$	-1 ± 6	$+72 \pm 39$	$+4 \pm 6$
20 s/AFOV	-43 ± 10	-2 ± 6	-34 ± 11	-3 ± 7
GF(20 s/AFOV)[1]	-25 ± 14	-2 ± 6	-17 ± 14	-3 ± 7
2D U-Net(20 s/AFOV)[2]	$+117 \pm 39$	$+1 \pm 4$	$+38 \pm 30$	$+2 \pm 6$
2.5D U-Net(20 s/AFOV)[3]	$+104 \pm 41$	0 ± 4	$+52 \pm 30$	$+1 \pm 5$
3D U-Net(20 s/AFOV)[4]	$+83 \pm 57$	0 ± 5	$+63 \pm 35$	$+5 \pm 6$
30 s/AFOV	-31 ± 12	0 ± 4	-22 ± 12	-1 ± 6
GF(30 s/AFOV)[1]	-15 ± 15	0 ± 4	-9 ± 14	-1 ± 5
2D U-Net(30 s/AFOV)[2]	$+71 \pm 42$	0 ± 3	$+22 \pm 20$	$+1 \pm 5$
2.5D U-Net(30 s/AFOV)[3]	$+101 \pm 41$	0 ± 3	$+42 \pm 26$	$+2 \pm 5$
3D U-Net(30 s/AFOV)[4]	$+72 \pm 48$	0 ± 3	$+45 \pm 31$	$+3 \pm 5$

[1] Set denoised through Gaussian smoothing.
[2] Set denoised through the 2D U-Net.
[3] Set denoised through the 2.5D U-Net.
[4] Set denoised through the 3D U-Net.

Table 5.3: 95% limits of agreement between the reference images for the 75 included lesions in the original (not denoised) and denoised (Gaussian filter, 2D U-Net, 2.5D U-Net and 3D U-Net) images for a given feature, for each scan duration.

Image set \ Measure	ΔSUV_{max} [g/mL]	ΔSUV_{mean} [g/mL]	ΔSUV_{SD} [g/mL]	ΔSUV_{peak} [g/mL]	ΔTLG [g]	ΔMTV [cm^3]
15 s/AFOV	[-1.21, 1.47]	[-0.82, 0.69]	[-0.22, 0.25]	[-0.67, 0.58]	[-7.47, 9.18]	[-2.66, 3.01]
GF(15 s/AFOV)[(1)]	[-1.99, 0.27]	[-1.08, 0.26]	[-0.46, 0.11]	[-1.02, 0.22]	[-7.67, 6.76]	[-2.41, 3.21]
2D U-Net(15 s/AFOV)[(2)]	[-2.40, 0.99]	[-1.22, 0.42]	[-0.48, 0.22]	[-1.36, 0.54]	[-10.68, 6.91]	[-2.91, 2.39]
2.5D U-Net(15 s/AFOV)[(3)]	[-2.05, 0.74]	[-1.24, 0.49]	[-0.38, 0.16]	[-1.17, 0.46]	[-8.60, 6.64]	[-2.29, 2.03]
3D U-Net(15 s/AFOV)[(4)]	[-2.25, 1.15]	[-1.24, 0.61]	[-0.44, 0.25]	[-1.07, 0.55]	[-7.83, 7.23]	[-2.20, 2.22]
20 s/AFOV	[-1.18, 1.33]	[-0.74, 0.67]	[-0.29, 0.32]	[-0.72, 0.66]	[-6.85, 7.49]	[-3.11, 3.16]
GF(20 s/AFOV)[(1)]	[-1.97, 0.61]	[-1.04, 0.40]	[-0.47, 0.19]	[-1.05, 0.40]	[-7.12, 5.86]	[-2.56, 3.04]
2D U-Net(20 s/AFOV)[(2)]	[-1.66, 1.03]	[-1.04, 0.65]	[-0.37, 0.25]	[-1.01, 0.63]	[-8.69, 7.23]	[-3.18, 2.74]
2.5D U-Net(20 s/AFOV)[(3)]	[-1.60, 0.74]	[-1.01, 0.62]	[-0.35, 0.21]	[-0.96, 0.50]	[-8.19, 6.13]	[-3.07, 2.51]
3D U-Net(20 s/AFOV)[(4)]	[-1.76, 1.03]	[-1.06, 0.68]	[-0.38, 0.26]	[-0.90, 0.66]	[-8.92, 11.00]	[-2.60, 3.05]
30 s/AFOV	[-0.92, 0.99]	[-0.70, 0.69]	[-0.19, 0.20]	[-0.58, 0.53]	[-5.87, 5.59]	[-2.28, 2.09]
GF(30 s/AFOV)[(1)]	[-1.38, 0.37]	[-0.82, 0.41]	[-0.33, 0.11]	[-0.77, 0.31]	[-6.23, 4.72]	[-2.10, 2.30]
2D U-Net(30 s/AFOV)[(2)]	[-1.55, 0.77]	[-0.94, 0.56]	[-0.31, 0.15]	[-0.92, 0.53]	[-7.84, 4.88]	[-2.31, 1.84]
2.5D U-Net(30 s/AFOV)[(3)]	[-1.26, 0.76]	[-0.85, 0.61]	[-0.25, 0.17]	[-0.79, 0.56]	[-5.60, 4.19]	[-1.94, 1.67]
3D U-Net(30 s/AFOV)[(4)]	[-1.39, 0.91]	[-0.79, 0.54]	[-0.29, 0.22]	[-0.72, 0.57]	[-5.60, 5.50]	[-1.67, 1.59]

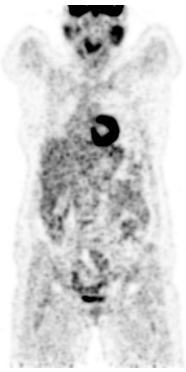
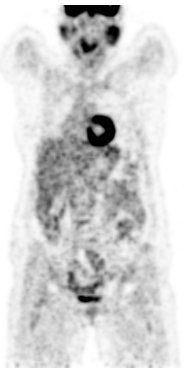
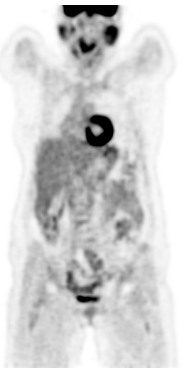
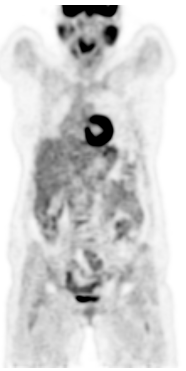
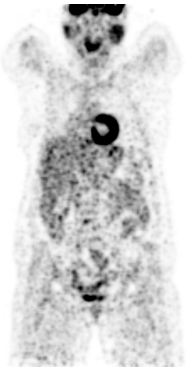
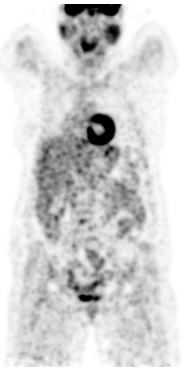
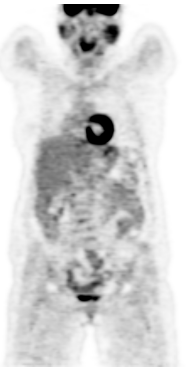
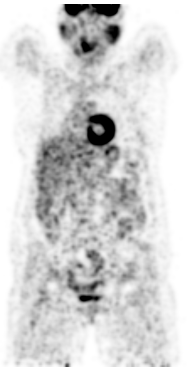
[(1)] Set denoised through Gaussian smoothing.
[(2)] Set denoised through the 2D U-Net.
[(3)] Set denoised through the 2.5D U-Net.
[(4)] Set denoised through the 3D U-Net.

Table 5.4: Median absolute deviation (1st, 3rd quartiles) between the reference images for the 75 included lesions in the original (not denoised) and denoised (Gaussian filter, 2D U-Net, 2.5D U-Net and 3D U-Net) images for a given feature, for each scan duration.

Image set \ Measure	$\lvert\Delta SUV_{max}\rvert$ [g/mL]	$\lvert\Delta SUV_{mean}\rvert$ [g/mL]	$\lvert\Delta SUV_{SD}\rvert$ [g/mL]	$\lvert\Delta SUV_{peak}\rvert$ [g/mL]	$\lvert\Delta TLG\rvert$ [g]	$\lvert\Delta MTV\rvert$ [cm^3]
15 s/AFOV	0.38 (0.18, 0.67)*	0.20 (0.08, 0.36)*	0.08 (0.05, 0.13)	0.21 (0.10, 0.34)*	1.23 (0.36, 2.94)	0.38 (0.19, 0.93)
GF(15 s/AFOV)[(1)]	0.89 (0.45, 1.31)	0.42 (0.18, 0.60)	0.14 (0.08, 0.27)	0.40 (0.15, 0.60)	1.45 (0.70, 3.35)	0.51 (0.19, 1.34)
2D U-Net(15 s/AFOV)[(2)]	0.83 (0.34, 1.25)	0.33 (0.16, 0.68)	0.13 (0.07, 0.24)	0.52 (0.24, 0.76)	1.65 (0.83, 3.22)	0.45 (0.26, 1.06)
2.5D U-Net(15 s/AFOV)[(3)]	0.68 (0.27, 1.17)	0.33 (0.10, 0.72)	0.12 (0.05, 0.21)	0.38 (0.25, 0.61)	1.37 (0.65, 3.22)	0.38 (0.19, 0.90)
3D U-Net(15 s/AFOV)[(4)]	0.79 (0.37, 1.22)	0.45 (0.16, 0.65)	0.15 (0.07, 0.25)	0.39 (0.23, 0.56)	1.47 (0.81, 2.90)	0.45 (0.26, 0.83)
Friedman test *p*-value	<0.001	0.002	<0.001	<0.001	0.061	0.460
20 s/AFOV	0.39 (0.19, 0.58)	0.22 (0.09, 0.38)*	0.07 (0.03, 0.14)	0.17 (0.07, 0.27)*	1.08 (0.45, 2.43)	0.38 (0.19, 0.96)
GF(20 s/AFOV)[(1)]	0.59 (0.26, 0.94)	0.27 (0.14, 0.53)	0.11 (0.05, 0.19)	0.28 (0.16, 0.51)	1.37 (0.64, 2.86)	0.90 (0.32, 1.22)
2D U-Net(20 s/AFOV)[(2)]	0.43 (0.20, 0.87)	0.24 (0.12, 0.46)	0.10 (0.03, 0.16)	0.27 (0.13, 0.43)	1.58 (0.70, 3.22)	0.58 (0.26, 1.38)
2.5D U-Net(20 s/AFOV)[(3)]	0.50 (0.26, 0.79)	0.27 (0.10, 0.47)	0.09 (0.04, 0.17)	0.32 (0.13, 0.45)	1.37 (0.64, 2.93)	0.51 (0.13, 1.15)
3D U-Net(20 s/AFOV)[(4)]	0.53 (0.21, 0.93)	0.29 (0.10, 0.57)	0.12 (0.06, 0.18)	0.29 (0.13, 0.44)	1.52 (0.59, 2.90)	0.70 (0.32, 1.18)
Friedman test *p*-value	0.013	0.029	0.005	<0.001	0.010	0.018
30 s/AFOV	0.23 (0.14, 0.41)	0.17 (0.09, 0.32)	0.05 (0.02, 0.09)	0.12 (0.05, 0.28)*	0.96 (0.46, 2.12)	0.45 (0.13, 0.83)
GF(30 s/AFOV)[(1)]	0.49 (0.25, 0.83)	0.22 (0.08, 0.39)	0.08 (0.04, 0.17)	0.19 (0.11, 0.41)	1.01 (0.34, 2.26)	0.45 (0.13, 0.83)
2D U-Net(30 s/AFOV)[(2)]	0.52 (0.23, 0.73)	0.26 (0.11, 0.40)	0.10 (0.04, 0.14)	0.26 (0.19, 0.41)	1.50 (0.51, 3.01)	0.58 (0.26, 1.09)
2.5D U-Net(30 s/AFOV)[(3)]	0.38 (0.18, 0.63)	0.19 (0.08, 0.36)	0.06 (0.04, 0.10)	0.22 (0.07, 0.34)	1.13 (0.35, 2.64)	0.38 (0.13, 0.86)
3D U-Net(30 s/AFOV)[(4)]	0.42 (0.19, 0.72)	0.20 (0.11, 0.37)	0.07 (0.03, 0.13)	0.19 (0.10, 0.33)	1.15 (0.52, 2.22)	0.38 (0.19, 0.83)
Friedman test *p*-value	<0.001	0.069	<0.001	<0.001	0.022	0.101

[(1)] Set denoised through Gaussian smoothing; [(2)] Set denoised through the 2D U-Net; [(3)] Set denoised through the 2.5D U-Net; [(4)] Set denoised through the 3D U-Net.
* Post Hoc analysis through the Wilcoxon signed-rank test between pairs of sets revealed statistically significant difference between every denoised set against the respective original (not denoised) set.

Table 5.5: Comparison of the different image sets — original and denoised through the Gaussian filter, the 2D U-Net, the 2.5D U-Net and the 3D U-Net — for the standard duration (70 s/AFOV) and the fast scans (15, 20 and 30 s/AFOV). The same coronal plane from a given patient of the test set is shown, for each image set, with the same intensity scale.

	Original	Gaussian filter	2D U-Net	2.5D U-Net	3D U-Net
70 s/AFOV					
30 s/AFOV					
20 s/AFOV					
15 s/AFOV					

6 | Discussion and Conclusions

In the present study, fast whole-body [^{18}F]FDG PET/CT scans were simulated with the aim of their denoising through deep-learning-based methods. The respective full-duration scans were the ones acquired following standard clinical protocol and were employed as the reference to the fast scans. The PET image reconstruction protocol fulfills EARL1 standards [16].

In PET/CT imaging, various factors influence image quality. These include scanner specifications (such as sensitivity and spatial resolution), reconstruction technique, imaging protocol and even patient habitus and demographics. In the context of this study, the most relevant are the activity administered to the patient and the acquisition time per bed position. The higher the administered activity, the more radioactive decays occur and, consequently, the more counts are detected by the scanner. In the same way, the longer time per bed, the more counts are collected by the detectors. The dataset employed in this work had an average injected activity of 3.4 ± 0.2 MBq/kg. This value is well within the range of the injected activity in the studies summarised in section 1.3, although being somewhat below average. The considered low-quality, noisy images corresponded to those acquired with 15, 20 and 30 seconds per bed position, which is significantly lower than those considered in the remaining studies. Thus, the noise level of the images in this work is considerably higher than that of the existing studies that aim to denoise PET/CT scans through deep-learning-based methods. However, let it be noted that for a given acquisition time, a PET/CT scanner with higher sensibility would output an image with a higher number of counts and, consequently, superior quality may be expected. This makes it harder to establish a clear comparison between two datasets if the respective scanners and imaging protocols are not the same.

An important feature of the present study, contrasting with the previous ones, is regarding the tumour's quantification methods. The semi-automatic segmentation performed individually for each set of images (through common algorithm-initialisation masks delineating the lesions) allowed an analysis in terms of metabolic tumour volume, besides the typical quantification concerning SUV measures. Furthermore, the optimisation of the benchmark method also constituted an uncommon approach that allowed for a fairer comparison between the developed deep-learning-based methods and the standard ones.

In the initial stage of development, a 2D U-Net was implemented as a first attempt to denoise the fast whole-body [^{18}F]FDG PET scans. This network's input consisted of the axial 2D images extracted from the original 3D volumes. Although promising, the resulting images showed artefacts from the 2D denoising, as is seen in table A.1 (appendix A). The voxel-wise

quantification analysis fell behind the 2.5D and 3D approaches. The regional quantification analysis revealed a fair performance of the 2D U-Net, compared to the other two networks, as it outperformed them in some instances. However, taking into consideration the above-mentioned perceptible artefacts in the output images, the 2D U-Net was considered a weak denoising method.

The second approach consisted of implementing a 2.5D U-Net, aiming to include, in the training of the model, the spatial information that was lacking in the 2D version. This network's input consisted of the axial, coronal and sagittal 2D planes extracted from the original 3D volumes. This strategy differs from the ones implemented by Xing et al. [10] and by Tsuchyia et al. [12], where the 2.5D networks' inputs were *N*-channel 2D images (of the same anatomical plane).

The 2.5D U-Net exhibits the overall best performance in the voxel-wise analysis (table 5.1). At the outset, the original fast scans had already shown reasonable agreement with the reference images, as per the low mean squared errors and high structural similarity index measures and intraclass correlation coefficients. Notwithstanding, further improvement was reported regarding these measures for the denoised sets, in terms of the noise level of the original images *versus* the DL-denoised images (as is seen in tables 5.5 and A.3).

Concerning healthy organ quantification, the denoising through the 2.5D U-Net resulted in a remarkable increase in signal-to-noise ratio in both the liver and lungs, while displaying no excessive differences in SUV_{mean}, relatively to the reference images. This indicates a significant decrease in the standard deviation of the intensity values in these expected-to-be uniform areas of radiopharmaceutical uptake. Even though an improvement in SNR is generally a good indicator of image quality, an excessively smoothed image would present a strong SNR while possibly having poor clinical potential, as it may have lost relevant information such as, for instance, the definition in the anatomical structures' or the lesions' borders.

Lesion quantification was performed in terms of the 95% limits of agreement between a given set of images and the respective reference (70 s/AFOV). These limits estimate the interval where 95% of the differences in a given lesion feature are expected to lie relatively to the reference. Thus, the narrower the interval is, the closer the measurements are. For the denoising through the 2.5D U-Net, the 95% limits of agreement concerning the lesion's SUV_{max} showed a variation that ranged from a decrease of about 2 g/mL to an increase of about 1 g/mL. This variation is not expected to have clinical impact, in most cases. Quantification in terms of absolute deviation from the reference lesions' SUV measures revealed a median variation of less than 1 g/mL and less than 0.50 cm^3 for the lesions' MTV. These results reinforce the previous statement.

Thereafter, a 3D implementation of the U-Net took place. The network's input consisted of a 3D patch randomly extracted from the original volume. This training strategy differs from the one by Mehranian et al. [11], where all the training patches were fed into the network in every epoch.

Although statistically significant differences were found in the voxel-wise analysis between the 3D and the 2.5D U-Nets (depicted in table 5.1), these discrepancies in MSE, SSIM and

ICC are in the order of the thousandths and are not expected to have clinical impact. This is supported by fig. A.4 (appendix A), where the voxel-wise analysis results of the denoising of the 20 s/AFOV set through the 2.5D and 3D U-Nets are plotted side-by-side, and by table 5.5, where the (visually very similar) outputs of the 2.5D and 3D U-Nets are also displayed side-by-side. Greater differences relatively to the reference images were observed for both the original and the benchmark (Gaussian filter) sets of images. However, the respective voxel-wise analysis still revealed strong agreement with the reference (MSE<0.1; SSIM and ICC>0.9). Even though this is the case, one can visually access the distinction in image quality between the different sets of images in table 5.5, once again.

The signal-to-noise ratio in regions of expected uptake uniformity in the liver and lungs demonstrated, once again, an improvement relatively to the reference images for deep-learning-based denoising. This positive variation in SNR is smaller for denoising through the 3D U-Net than through the 2.5D U-Net. In the lungs, it is slightly greater for the denoising through the 3D U-Net, still the improvement is similar overall, as one can see through the images in table 5.5.

Regarding lesion quantification, the 3D U-Net had a similar performance to the 2.5D U-Net, with feature variations rendered unimportant. The greatest deviations to the reference in the lesion quantification analysis of the original and denoised sets were consistently observed for the 15-s/AFOV-based sets, as expected. The median absolute deviations from the reference images' lesion features are reported in table 5.4. These results reveal small variations, once again, not expected to have clinical impact.

In some instances, the smaller deviations in the lesions' features were observed for the original (not denoised) sets, with statistically significant differences (p<0.05), compared to the remaining sets of images. This means that the alteration in some features was exacerbated in the denoising. However, the resulting deviations, although greater than those observed pre-denoising, were still considered clinically insignificant. Even though this is the case, the smoothing observed in the DL-denoised images may make it harder for the clinicians to discern the lesions from their surroundings.

Summing up, the deep-learning-based denoising models for fast whole-body [^{18}F]FDG PET/CT scans developed in this study proved to have potential to achieve images with clinically-suitable quantitative parameters. Deep-learning-based denoising outperformed optimised Gaussian filter in every instance. The 20 s/AFOV scans with post-processing with the 2.5D U-Net or the 3D U-Net seemed to be the best compromise between scan duration and image quality, compared to the 15 and 30-s/AFOV-based scans.

The goals outlined in section 1.2 were reached. The implementation of deep-learning-based methods for the denoising of fast PET/CT scans was successfully conducted and exhibited promising results in the quantitative analysis performed.

6.1 Limitations

This study's limitations, summarised below, must be considered:

1. In this study, 92 whole-body [^{18}F]FDG PET/CT scans were used to train each of the three implemented networks, which is short of ideal. When training a deep learning algorithm there is the underlying risk of employing a dataset of insufficient size, which happens often when dealing with medical data. For the network to properly learn how to perform a given task, its training inputs must represent, as much as possible, all real-world inputs. Thus, if too few examples are provided in the training of the network, the resulting model may not generalise well, when presented to new data.

2. The images of the dataset were acquired with the same PET/CT scanner. This means that the samples used to train and test the network did not contain inter-scanner variability. As the image quality varies with the scanner, this constitutes another limiting factor for network generalisation.

3. Besides the lack of inter-scanner variability in the dataset, it is also important to regard that the images included followed EARL1 standards [16]. Therefore, the networks trained in this study are not expected to perform as well on images that follow EARL2 standards [30], for instance.

4. The fast scans that served as the networks' input were obtained by cropping the raw data of the standard acquisition to the desired duration and performing the reconstruction anew. These 15, 20 and 30 seconds time-frames were extracted from the centre of the full-duration window. This means that the fast whole-body PET/CT images employed were simulated. In reality, if the PET data corresponded to real-time short acquisitions, the radiopharmaceutical movement, as well as other parameters such as the evolution of radioactive decay, would be distinct. However, if the fasts scans were real, the respective reference (high-quality) images, necessary for supervised learning, would not match. The pairs would be acquired independently, which would result in different patient positioning and, consequently, physical point mismatch. Moreover, as was brought up before, the radiopharmaceutical anatomical dispersion may be considerably disparate. This, along with the impracticality of the alternative, led us to opt for the so-called simulated fast scans, given this study's task.

5. Still on the matter of image acquisition, another aspect to consider is that extra-long acquisition duration scans would be the ideal reference images, rather than the standard duration scans. By doing so, the resulting high-quality images would constitute an improved target in the training. However, to not interfere with clinical protocol and bearing in mind the patient's comfort, the standard 70 s/AFOV scans were utilised.

6. Lastly, regarding the models' performance evaluation, a clinical qualitative analysis of the denoised fast scans is lacking. This is especially important to assess if the denoised

fast PET/CT images keep their clinical value for diagnostic, staging and prognostic. Thus, an evaluation of the images by experienced clinicians would be crucial.

6.2 Future Work

Given the limitations described above, it is undeniable that there is space for improvement in the work to be carried from here on:

1. Firstly, the collection of a larger dataset (both in terms of size and variability) is a priority. Images acquired through different PET/CT scanners and clinical protocols that meet different standards would improve the model's generalisation.

2. Besides the Gaussian filter, other denoising methods, such as non-local means denoising, could serve as additional benchmarks. This would only strengthen the study and further conclude on the advantages of deep-learning-based denoising.

3. Ideally, the deep-learning-based model should be able to remove any level of noise from the input images. To achieve this, the network must be trained with PET/CT images different levels of noise, rather than training one network per "noise category", i.e., for a specific scan duration. Accordingly, a possible future strategy would be to encompass in the training set all the different scan durations.

4. Still on the matter of improving the network's training, longer-than-standard-acquisition-time scans could provide better quality targets, as was discussed before.

5. The development/implementation and training of deep learning algorithms for the denoising of PET/CT images of other radiopharmaceuticals (e.g. ^{68}Ga-labelled PSMA) or specific body parts (e.g. lung, cerebral or cardiac PET/CT) would bring similar benefits to biomedical imaging as this study aimed to bring.

6. Last but not least, the qualitative evaluation of the denoised images by experienced physicians would constitute an exponential improvement in the study's relevance and robustness. The deep-learning-based denoising models were developed with the intention of future clinical implementation. Therefore, the most important measure to assess their viability is the clinicians' approval.

Bibliography

[1] H. Sung, J. Ferlay, R. L. Siegel, *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, pp. 209–249, 3 2021-05. DOI: 10.3322/CAAC.21660.

[2] S. P. Rowe and M. G. Pomper, "Molecular imaging in oncology: Current impact and future directions," *CA: A Cancer Journal for Clinicians*, vol. 72, 4 2022. DOI: 10.3322/caac.21713.

[3] Y. F. Tai and P. Piccini, "Applications of positron emission tomography (PET) in neurology," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 75, 5 2004. DOI: 10.1136/jnnp.2003.028175.

[4] B. X. Tran, G. T. Vu, G. H. Ha, *et al.*, "Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study," *Journal of Clinical Medicine*, vol. 8, 3 2019. DOI: 10.3390/jcm8030360.

[5] N. Aide, C. Lasnon, C. Desmonts, I. S. Armstrong, M. D. Walker, and D. R. McGowan, "Advances in PET/CT Technology: An Update," *Seminars in Nuclear Medicine*, vol. 52, pp. 286–301, 3 2022, Advancement in Instrumentation for Molecular Imaging. DOI: https://doi.org/10.1053/j.semnuclmed.2021.10.005.

[6] K. Weyts, C. Lasnon, R. Ciappuccini, *et al.*, "Artificial intelligence-based PET denoising could allow a two-fold reduction in [18F]FDG PET acquisition time in digital PET/CT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, pp. 3750–3760, 11 2022. DOI: 10.1007/s00259-022-05800-1.

[7] G. Bonardel, A. Dupont, P. Decazes, *et al.*, "Clinical and phantom validation of a deep learning based denoising algorithm for F-18-FDG PET images from lower detection counting in comparison with the standard acquisition," *EJNMMI Physics*, vol. 9, p. 36, 1 2022. DOI: 10.1186/s40658-022-00465-z.

[8] K. Katsari, D. Penna, V. Arena, *et al.*, "Artificial intelligence for reduced dose 18F-FDG PET examinations: a real-world deployment through a standardized framework and business case assessment," *EJNMMI Physics*, vol. 8, 1 2021. DOI: 10.1186/s40658-021-00374-7.

[9] A. S. Chaudhari, E. Mittra, G. A. Davidzon, *et al.*, "Low-count whole-body PET with deep learning in a multicenter and externally validated study," *npj Digital Medicine*, vol. 4, 1 2021. DOI: 10.1038/s41746-021-00497-2.

[10] Y. Xing, W. Qiao, T. Wang, *et al.*, "Deep learning-assisted PET imaging achieves fast scan/low-dose examination," *EJNMMI Physics*, vol. 9, 1 2022. DOI: 10.1186/s40658-022-00431-9.

[11] A. Mehranian, S. D. Wollenweber, M. D. Walker, *et al.*, "Image enhancement of whole-body oncology [18F]-FDG PET scans using deep neural networks to reduce noise," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, 2 2022. DOI: 10.1007/s00259-021-05478-x.

[12] J. Tsuchiya, K. Yokoyama, K. Yamagiwa, *et al.*, "Deep learning-based image quality improvement of 18F-fluorodeoxyglucose positron emission tomography: a retrospective observational study," *EJNMMI Physics*, vol. 8, 1 2021. DOI: 10.1186/s40658-021-00377-4.

[13] D. W. Townsend, "Physical principles and technology of clinical PET imaging," *Annals of the Academy of Medicine Singapore*, vol. 33, 2 2004.

[14] S. Vandenberghe, E. Mikhaylova, E. D'Hoe, P. Mollet, and J. S. Karp, "Recent developments in time-of-flight PET," *EJNMMI Physics*, vol. 3, 1 2016. DOI: 10.1186/s40658-016-0138-3.

[15] H. A. Ziessman, J. P. O'Malley, J. H. Thrall, and F. H. Fahey, *Nuclear Medicine: Fourth Edition*. 2013. DOI: 10.1016/C2010-0-65632-X.

[16] R. Boellaard, M. J. O'Doherty, W. A. Weber, *et al.*, "FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: Version 1.0," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 37, 1 2010. DOI: 10.1007/s00259-009-1297-4.

[17] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, 2020. DOI: 10.1016/j.bspc.2020.102036.

[18] C. P. Behrenbruch, S Petroudi, S Bond, J. D. Declerck, F. J. Leong, and J. M. Brady, "Image filtering techniques for medical image post-processing: An overview," *The British Journal of Radiology*, vol. 77, S126–S132, suppl2 2004. DOI: 10.1259/bjr/17464219.

[19] K. Matsubara, M. Ibaraki, M. Nemoto, H. Watabe, and Y. Kimura, "A review on AI in PET imaging," *Annals of Nuclear Medicine*, vol. 36, 2 2022. DOI: 10.1007/s12149-021-01710-8.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org (visited on 2023-04-11).

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 7553 2015-05. DOI: 10.1038/nature14539.

[22] J. Teuwen and N. Moriakov, "Chapter 20 - Convolutional neural networks," in S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. Academic Press, 2020, pp. 481–501. DOI: 10.1016/B978-0-12-816176-0.00025-9.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," vol. 9351, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," IEEE, 2015-12, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015. DOI: 10.48550/arXiv.1412.6980.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, 4 2004. DOI: 10.1109/TIP.2003.819861.

[27] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, 2 1979. DOI: 10.1037/0033-2909.86.2.420.

[28] C. S. Constantino, F. P. M. Oliveira, M. Silva, *et al.*, "Are lesion features reproducible between 18F-FDG PET/CT images when acquired on analog or digital PET/CT scanners?" *European Radiology*, vol. 31, pp. 3071–3079, 5 2021. DOI: 10.1007/s00330-020-07390-8.

[29] C. S. Constantino, S. Leocádio, F. P. M. Oliveira, *et al.*, "Evaluation of Semiautomatic and Deep Learning–Based Fully Automatic Segmentation Methods on [18F]FDG PET/CT Images from Patients with Lymphoma: Influence on Tumor Characterization," *Journal of Digital Imaging*, 2023. DOI: 10.1007/s10278-023-00823-y.

[30] R. Boellaard, R. Delgado-Bolton, W. J. Oyen, *et al.*, "FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 42, 2 2015. DOI: 10.1007/s00259-014-2961-x.

A | Additional Results

Below, figs. A.1, A.2 and A.3 display the box and parallel coordinates plots for MSE, SSIM and ICC, for the three sets of images considered in the voxel-wise analysis – original (not denoised) and denoised through the Gaussian filter and the 2D U-Net –, for each scan duration.

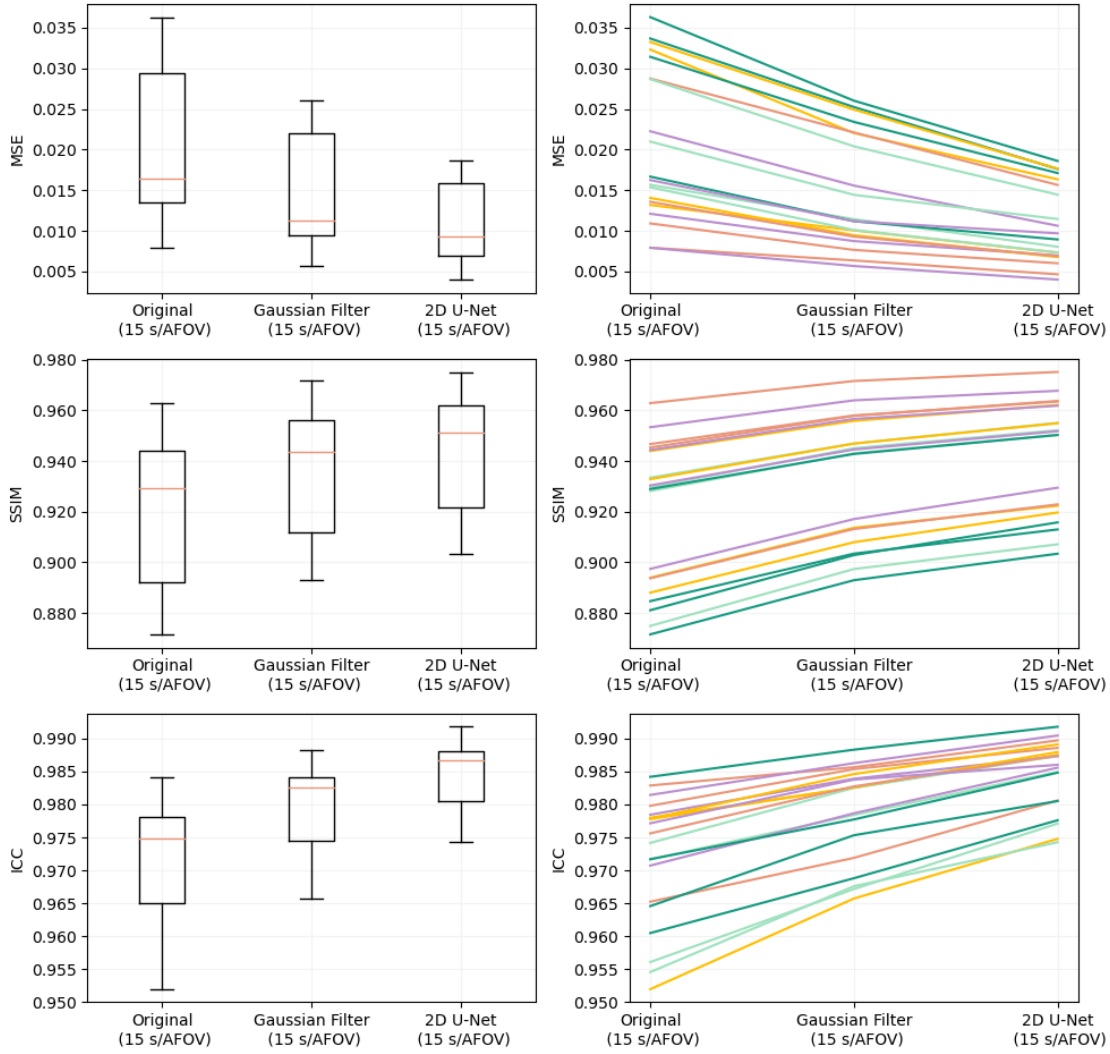


Figure A.1: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 15 s/AFOV images, the GF-denoised images and the DL-denoised images (2D U-Net).

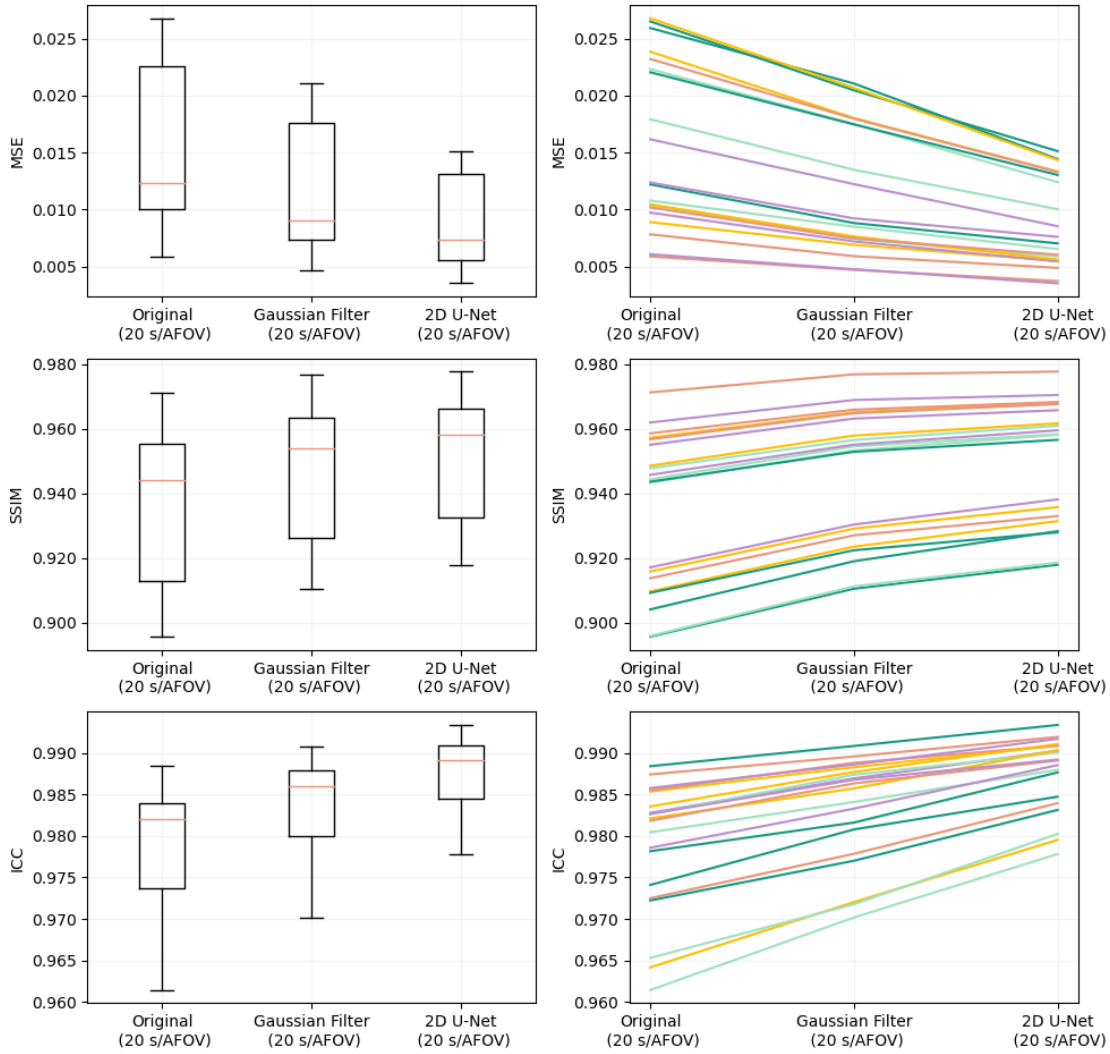


Figure A.2: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 20 s/AFOV images, the GF-denoised images and the DL-denoised images (2D U-Net).

These plots aid the interpretation of the results shown in table 5.1, of the voxel-wise analysis performed on the sets denoised through the 2D U-Nets, in contrast with the original and GF-denoised image sets, for each scan duration.

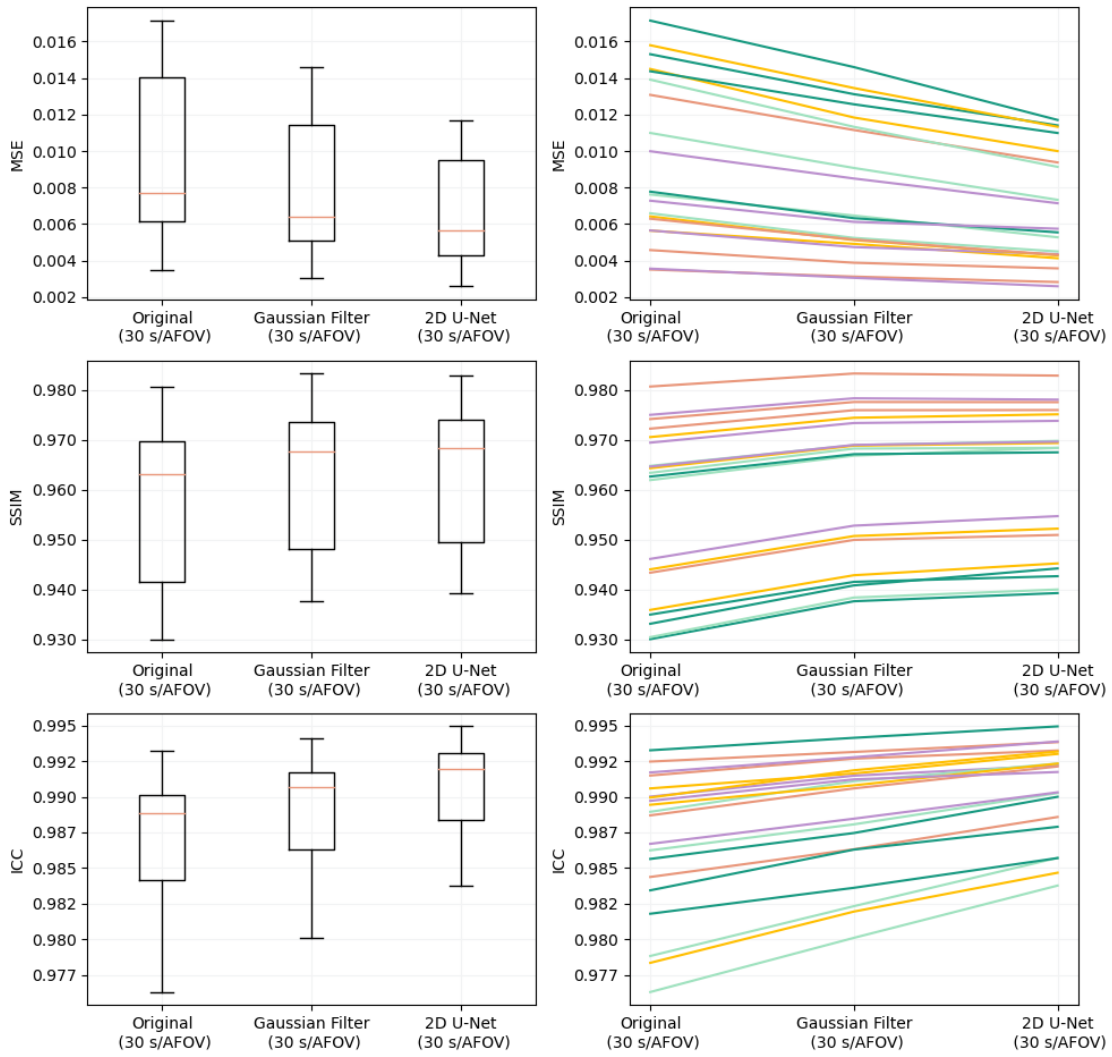


Figure A.3: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 30 s/AFOV images, the GF-denoised images and the DL-denoised images (2D U-Net).

Although the quantitative analysis of the images denoised through the 2D U-Net revealed fair performance, as reported in section 5.2, the resulting images showed artefacts that arose from the "axial" denoising. Table A.1 contains a sagittal and a coronal plane where it is possible to identify the presence of "horizontal blurring", given the respective reference standing alongside.

Table A.1: Presence of artefacts that arose from the 2D axial denoising in the sagittal and coronal planes from a patient of the test set, in the set denoised through the 2D U-Net and the respective reference. All images are displayed in the same intensity scale.

	2D U-Net(15 s/AFOV)	Reference (70 s/AFOV)
Sagittal		
Coronal		

Fig. A.4 displays the box and parallel coordinates plots for MSE, SSIM and ICC, for four of the 20-s/AFOV-based sets of images – original (not denoised) and denoised through the Gaussian filter, the 2.5D U-Net and the 3D U-Net.

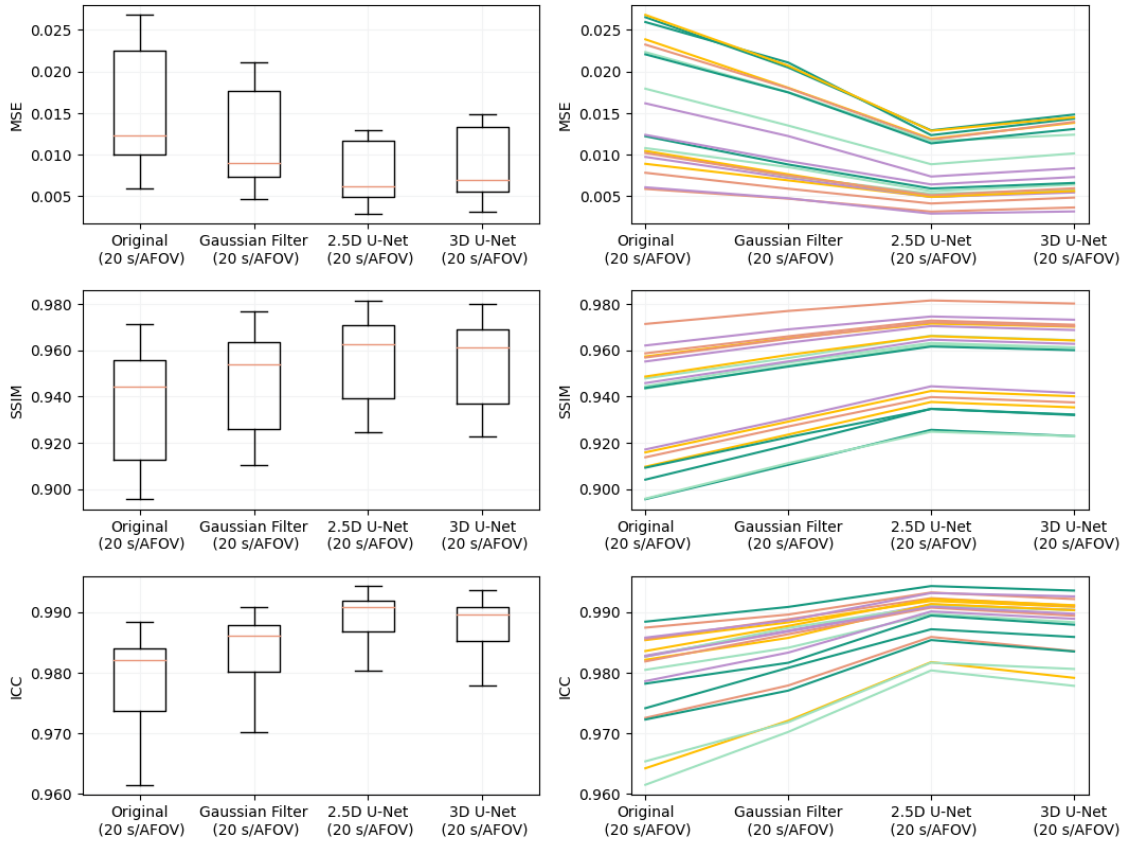


Figure A.4: Box and parallel coordinates plots of MSE, SSIM and ICC relative to the reference images, from within the test set, for the original (not denoised) 30 s/AFOV images, the GF-denoised images and the DL-denoised images by both the 2.5D U-Net and the 3D U-Net.

These plots were included to establish, more closely, the comparison between the voxel-wise quantification of the denoising through the 2.5D and the 3D U-Nets. The results are exhibited for the 20-s/AFOV-based sets, as this is the intermediate scan duration considered in this study.

Table A.2 displays the average relative difference, expressed as a percentage relatively to the reference, of the included lesions' features in the original, GF-denoised and DL-denoised sets of images, for the three scan durations.

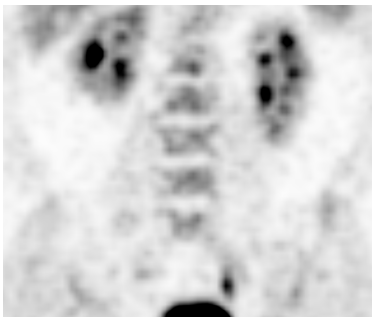
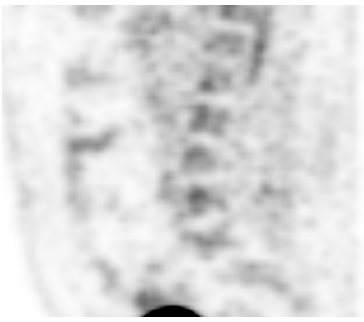
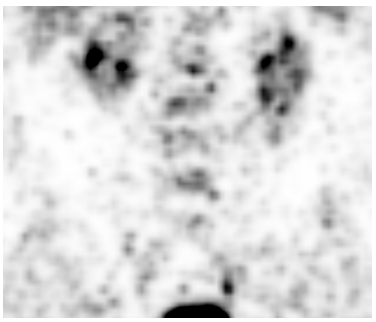
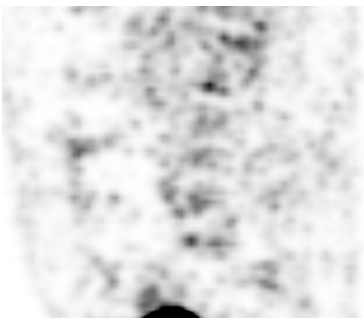
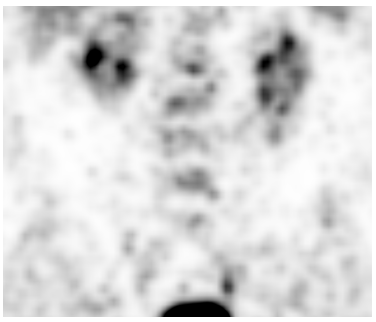
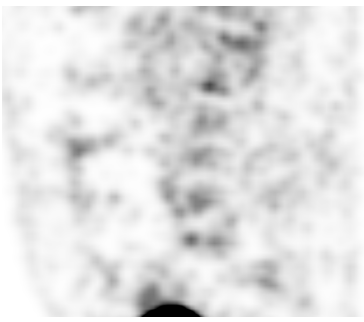
Table A.2: Average relative difference to the reference images (± standard deviation) for the 75 included lesions in the original (not denoised) and denoised (Gaussian filter, 2D U-Net, 2.5D U-Net and 3D U-Net) images for a given feature, for each scan duration.

Image set \ Measure	ΔSUV_{max} [%]	ΔSUV_{mean} [%]	ΔSUV_{SD} [%]	ΔSUV_{peak} [%]	ΔTLG [%]	ΔMTV [%]
15 s/AFOV	$+1 \pm 12$	-2 ± 11	$+4 \pm 22$	-2 ± 9	$+8 \pm 37$	$+14 \pm 49$
GF(15 s/AFOV)[(1)]	-17 ± 11	-13 ± 11	-25 ± 18	-12 ± 9	-1 ± 34	$+17 \pm 48$
2D U-Net(15 s/AFOV)[(2)]	-18 ± 19	-14 ± 16	-27 ± 30	-15 ± 14	-16 ± 34	$+5 \pm 64$
2.5D U-Net(15 s/AFOV)[(3)]	-17 ± 17	-14 ± 15	-24 ± 26	-4 ± 14	-12 ± 31	$+6 \pm 49$
3D U-Net(15 s/AFOV)[(4)]	-16 ± 19	-12 ± 17	-25 ± 29	-11 ± 14	-8 ± 31	$+8 \pm 43$
20 s/AFOV	$+2 \pm 12$	-1 ± 11	$+5 \pm 22$	-1 ± 9	$+7 \pm 38$	$+11 \pm 49$
GF(20 s/AFOV)[(1)]	-13 ± 11	-10 ± 11	-18 ± 17	-9 ± 9	$+1 \pm 36$	$+15 \pm 51$
2D U-Net(20 s/AFOV)[(2)]	-8 ± 15	-7 ± 15	-13 ± 25	-8 ± 11	-6 ± 35	$+6 \pm 52$
2.5D U-Net(20 s/AFOV)[(3)]	-11 ± 14	-7 ± 14	-15 ± 23	-9 ± 11	-9 ± 34	$+2 \pm 52$
3D U-Net(20 s/AFOV)[(4)]	-10 ± 18	-7 ± 16	-16 ± 27	-6 ± 13	$+1 \pm 32$	$+14 \pm 49$
30 s/AFOV	0 ± 9	0 ± 10	$+1 \pm 17$	-1 ± 7	$+2 \pm 31$	$+6 \pm 47$
GF(30 s/AFOV)[(1)]	-10 ± 8	-6 ± 10	-15 ± 14	-7 ± 6	-1 ± 32	$+9 \pm 49$
2D U-Net(30 s/AFOV)[(2)]	-10 ± 11	-7 ± 12	-16 ± 19	-7 ± 8	-8 ± 33	$+2 \pm 48$
2.5D U-Net(30 s/AFOV)[(3)]	-7 ± 11	-5 ± 12	-11 ± 18	-5 ± 8	-4 ± 32	$+4 \pm 51$
3D U-Net(30 s/AFOV)[(4)]	-7 ± 13	-5 ± 11	-10 ± 23	-4 ± 9	$+1 \pm 33$	$+9 \pm 46$

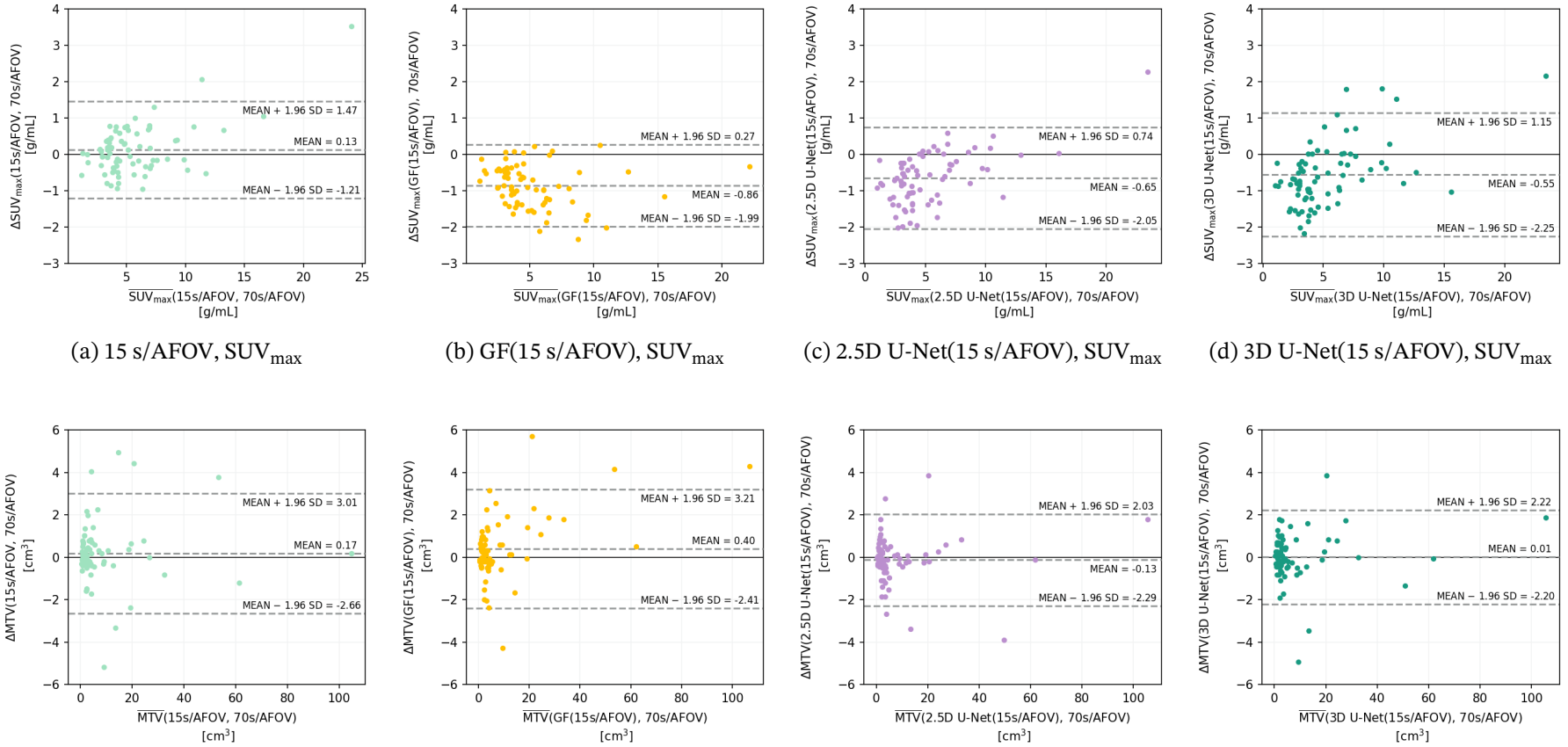
[(1)] Set denoised through Gaussian smoothing. [(2)] Set denoised through the 2D U-Net. [(3)] Set denoised through the 2.5D U-Net. [(4)] Set denoised through the 3D U-Net.

Table A.3 was included to better display the denoising through the 2.5D U-Net, in contrast to the other image sets (reference, original and benchmark). A close-up of a region in the body is shown, to aid the visual perception of image quality.

Table A.3: Close-up of a region in the sagittal and coronal planes from a patient of the test set. Corresponding planes in the original 15 s/AFOV set, the GF-denoised set, the set denoised through the 2.5D U-Net and the respective reference. All images are displayed in the same intensity scale.

Image set	Coronal	Sagittal
70 s/AFOV		
15 s/AFOV		
GF(15 s/AFOV)		
2.5D U-Net(15 s/AFOV)		

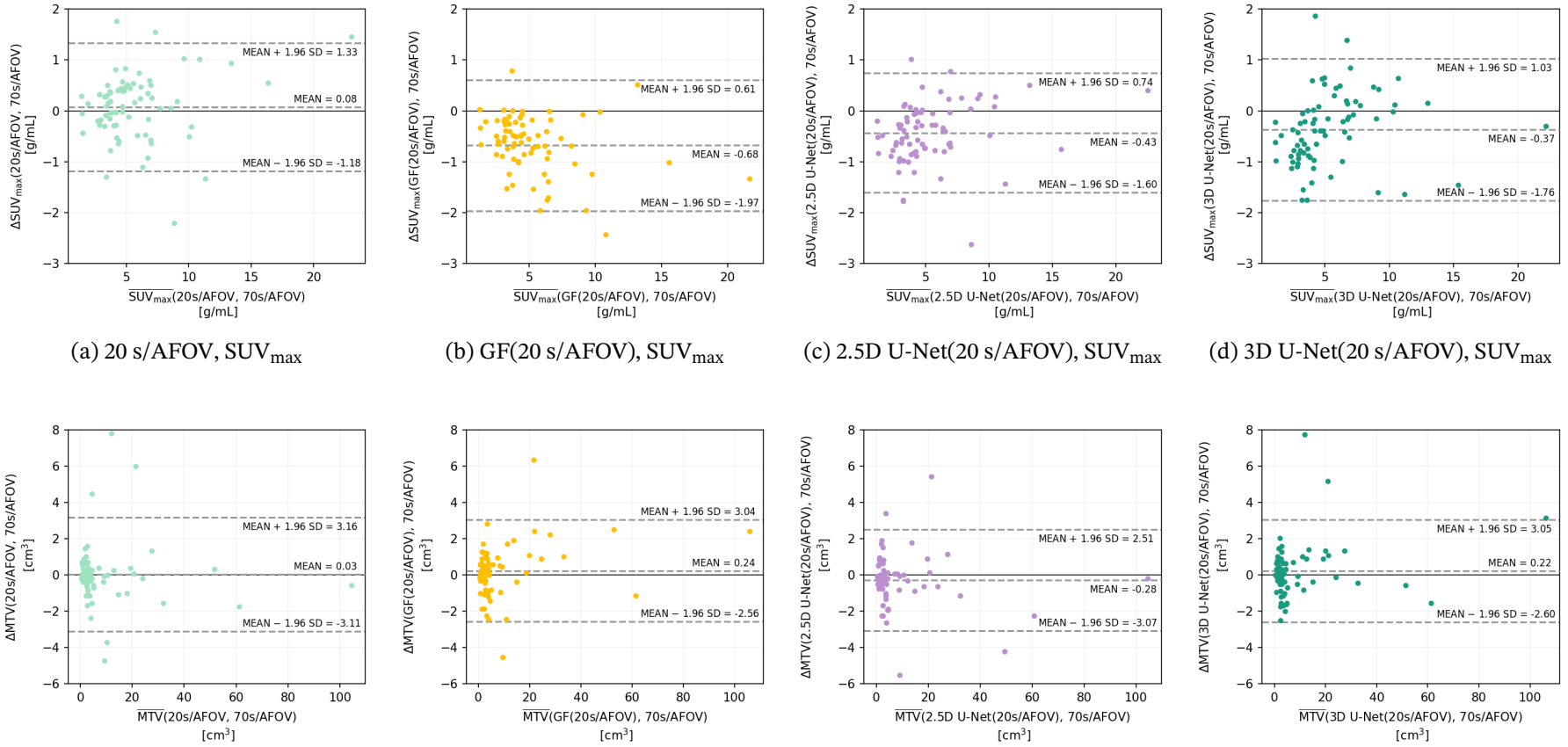
Lastly, figs. A.5, A.6 and A.7 show the Bland-Altman plots regarding SUV_{max} and MTV in the included lesions of the different image sets, for the three scan durations. These plots aid the interpretation of table 5.3.



(a) 15 s/AFOV, SUV_{max} (b) GF(15 s/AFOV), SUV_{max} (c) 2.5D U-Net(15 s/AFOV), SUV_{max} (d) 3D U-Net(15 s/AFOV), SUV_{max}

(e) 15 s/AFOV, MTV (f) GF(15 s/AFOV), MTV (g) 2.5D U-Net(15 s/AFOV), MTV (h) 3D U-Net(15 s/AFOV), MTV

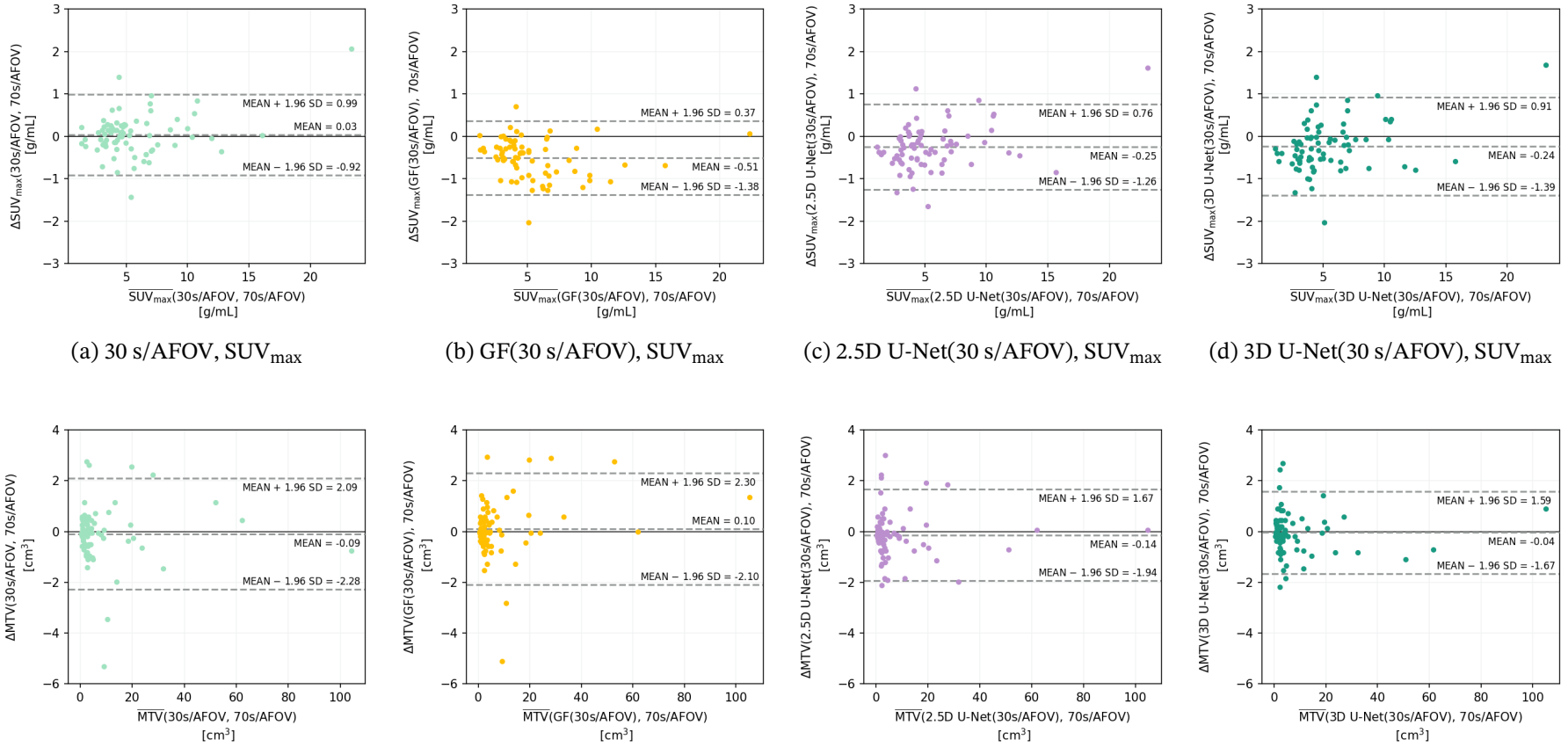
Figure A.5: Bland-Altman plots regarding SUV_{max} and MTV of the 75 included lesions, for no denoising, denoising through the Gaussian filter and denoising through the 2.5D and 3D U-Nets, for 15 s/AFOV.



(a) 20 s/AFOV, SUV_{max} (b) GF(20 s/AFOV), SUV_{max} (c) 2.5D U-Net(20 s/AFOV), SUV_{max} (d) 3D U-Net(20 s/AFOV), SUV_{max}

(e) 20 s/AFOV, MTV (f) GF(20 s/AFOV), MTV (g) 2.5D U-Net(20 s/AFOV), MTV (h) 3D U-Net(20 s/AFOV), MTV

Figure A.6: Bland-Altman plots regarding SUV_{max} and MTV of the 75 included lesions, for no denoising, denoising through the Gaussian filter and denoising through the 2.5D and 3D U-Nets, for 20 s/AFOV.



(a) 30 s/AFOV, SUV_{max} (b) GF(30 s/AFOV), SUV_{max} (c) 2.5D U-Net(30 s/AFOV), SUV_{max} (d) 3D U-Net(30 s/AFOV), SUV_{max}

(e) 30 s/AFOV, MTV (f) GF(30 s/AFOV), MTV (g) 2.5D U-Net(30 s/AFOV), MTV (h) 3D U-Net(30 s/AFOV), MTV

Figure A.7: Bland-Altman plots regarding SUV_{max} and MTV of the 75 included lesions, for no denoising, denoising through the Gaussian filter and denoising through the 2.5D and 3D U-Nets, for 30 s/AFOV.