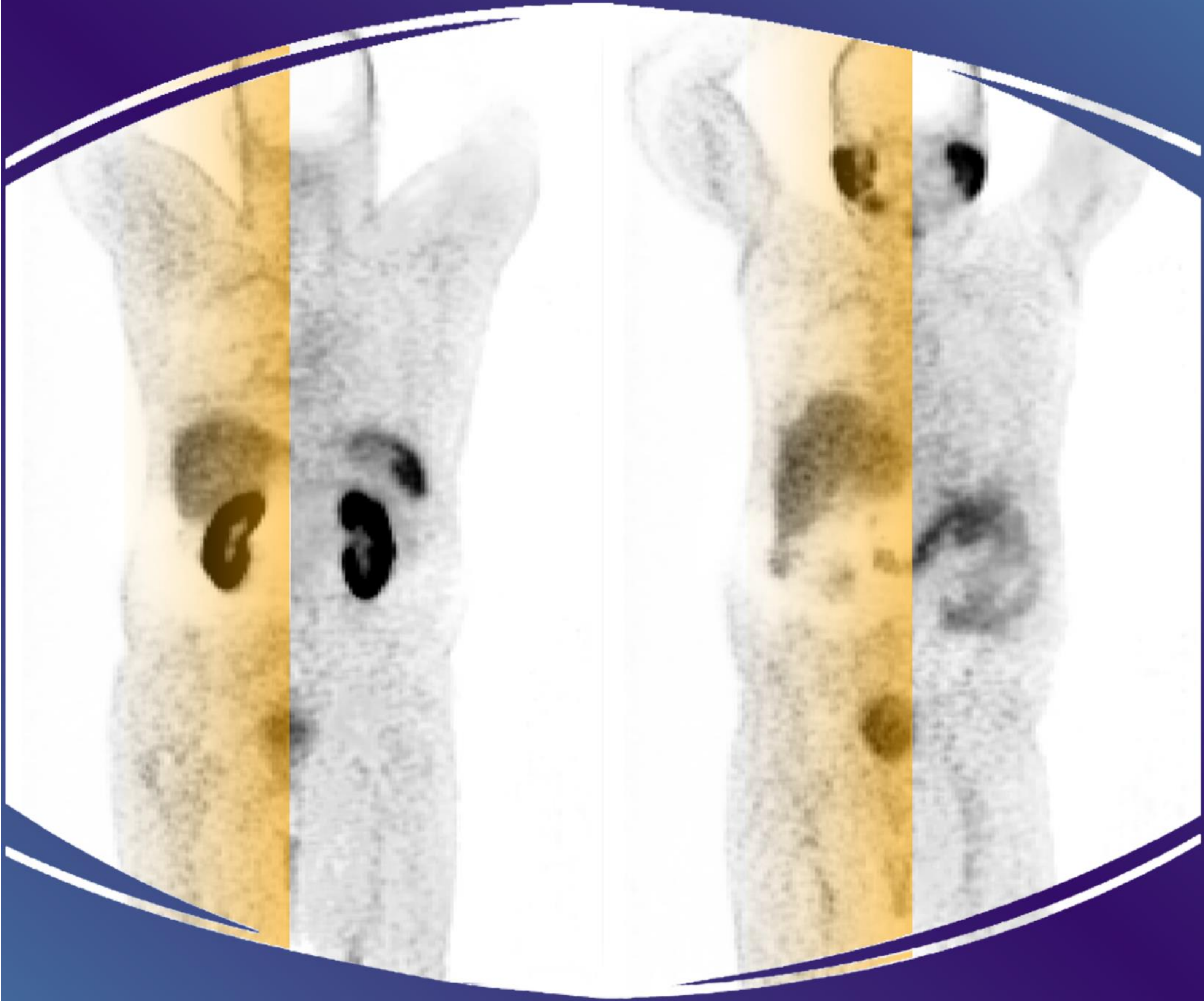
Deep Learning-Based PET Image Correction Toward Quantitative Imaging



Zohreh Shahpouri

460145

Data Science for life Science

June 01, 2024

Dr. Isaac Shiri

INSELGRUPPE

[1]

Deep Learning-Based PET Image Correction Toward Quantitative Imaging

Author

Zohreh Shahpouri

Student number

460145

Study

Data Science for Life Science

Institute

Hanze University of Applied Sciences

Institute of Life Science & Technology

Supervisor

Dr. Isaac Shiri

Abstract

Recent advancements in deep learning (DL) offer significant advantages in PET imaging, particularly in enhancing attenuation scatter correction (ASC) and artifact removal. However, practical implementation remains challenging due to variability in scanner types and radiotracer distributions. We aim to develop an Integrated Multi-Center DL Model (IMCM) to address the direct ASC of PET images and evaluate its performance in removing image artifacts.

A total of 270 clean and artifact-free images were selected from a collection of over 2000 patient databases undergoing ^{68}Ga and ^{18}F-FDG PET/CT scans across seven centers. Additionally, a collection of artifact-containing data was also gathered to assess the model's performance in artifact removal. Three data sets were designated for external testing: Cross-Center, Cross-Tracer, and Artifact-Correction evaluation (on ^{68}Ga, ^{18}F-FDG, and artifactual data set, respectively). A dedicated 3D-UNet model employing a deep supervision strategy was trained on artifact-free images from four centers. The model's performance was then evaluated voxel-wise intensity, quantitatively and qualitatively for ASC and artifact correction on the external test sets.

For the internal centers, the IMCM model achieved a Mean Error (ME) of -0.56±0.74, a Mean Absolute Error (MAE) of 1.28±0.37, and a Structural Similarity Index (SSIM) of 0.93±0.03. For the external test sets, IMCM yielded an ME of -1.92±0.58 and -0.54±0.13, an MAE of 2.38±0.76 and 0.69±0.12, and an SSIM of 0.89±0.03 and 0.78±0.10 for the Cross-Center and Cross-Tracer, respectively. IMCM successfully corrected motion and halo artifacts in both ^{68}Ga data and ^{18}F-FDG images.

The developed model effectively addressed variations in scanner types and radiotracers, demonstrating its adaptability, generalizability, and effectiveness in different clinical scenarios for direct ASC and artifact correction. This study shows the potential of DL to provide accurate, artifact-free PET images, offering a promising alternative to CT-based ASC.

Keywords: Deep learning, attenuation scatter correction, CT-less PET.

Abbreviation

Positron Emission Tomography	PET
Magnetic Resonance Imaging	MRI
Attenuation correction	AC
Attenuation and scatter correction	ASC
Computed Tomography	CT
Gallium-68	^{68}Ga
Fluorodeoxyglucose	FDG
Non attenuation scatter correction	NAC
CT based attenuation-scatter correction	MAC
Time of flight	TOF
Maximum likelihood estimation of activity and attenuation	MLAA
Long axial field of view	LAFOV
Artificial Intelligent	AI
Federated learning	FL
Prostate-Specific Membrane Antigen	PSMA
Anatomy-dependent correction model	ADCM
Deep learning model-based attenuation correction	DL
Standard uptake value	SUV
Integrated multi-Center model	IMCM
Tuned Transfer Learning for IMCM model	TL-MC

Table of Contents

Introduction

Positron Emission Tomography (PET) is the gold standard of molecular imaging modalities for the non-invasive study of various diseases (1–3). Numerous patients undergo PET scans worldwide for staging and restaging cancer, evaluating treatment diagnostic, radiation therapy planning, diagnosing neurological disorders, assessing myocardial perfusion, and surgical planning (4–6).

During a whole-body PET image creation, more than 50% of all recorded photons are resulted from at least one Compton scatter fraction before capturing by detectors (7–9). Photon scattering occurs due to dense materials in the patient's body and surrounding area, which causes energy loss(7) . A misplaced line of response (LOR) is formed by a scattered, attenuated photon, which has not been filtered out after energy window discrimination and random coincidence correction technique (10). So, scatter and attenuation phenomena lead to miscalculation of radiopharmaceutical distribution inside the body (7,10). Attenuation and Scatter correction (ASC) has a critical role in achieving a high-quality image interpretation and acceptable quantitative analysis of PET scans (Figure 1) (11,12).

ASC was initially performed using transmission scans with external radioactive sources to measure photon attenuation through the body (13). These measurements were used to create attenuation coefficient maps (μ-maps), which are essential for correcting PET images for the effects of photon attenuation and scatter (14,15). In the early 2000s, by introducing PET/CT technology, ASC began to use a CT scanner for modeling μ-maps (16–18). These μ-maps generated from CT scans provide precise anatomical details, significantly improving the accuracy of attenuation correction by differentiating between various tissues (14,15).

While various research has been done to create μ-maps from proton density information, ASC has remained a challenge in Magnetic Resonance Imaging (MRI) based AC (19–24). Despite the implementation of CT or MRI for ASC, artifacts, which are anomalies in the final images and do not correspond to the authentic radiotracer distribution within the body, can still occur (25–28). Patient motion during or between two scans complicates the alignment of PET with CT or MR images, causes mismatch, misregistration, or motion artifacts (25,26,29,30). Moreover, in reconstruction with ASC, neighboring areas to high-activity organs, such as the kidney bladder, might be assigned negative or zero values, leading to halo artifacts in clinical observations (31,32). Halo artifacts are very common in ^{68}Ga-PET imaging, which is widely used in prostate and pelvic cancer diagnosis, staging, and treatment planning (33–37). This artifact might change the interpretation of clinical diagnosis (33,38). To prevent these artifacts, giving diuretics often makes the patient more uncomfortable and increases the chance of motion artifacts, which makes the image quality and readability even worse (39–41). The presence of artifacts can significantly decrease the image quality and accuracy of interpretation and result in misdiagnoses. Consequently, even repeating scans fails to resolve the issue and can lead to an increased cumulative total body dose, and longer waiting times (39–41).

Many algorithms have been proposed for generating μ-maps, such as the maximum likelihood estimation of activity and attenuation, popularly known as MLAA (42) estimates radiotracer distribution and the attenuation map simultaneously from the emission data (42). However, it has been limited by poor coincidence time resolution which affects positional uncertainty. Later, with the introduction of time of flight (TOF) and spatial time revolution, the accuracy and reconstruction convergence in MLAA and similar algorithms are significantly enhanced (43–45). Whole-body PET scanners with a long axial field of view (LAFOV) have significantly improved quantification and image resolution (8,9,18,46–48). There has been tremendous improvement regarding ASC and the incorporation of advanced technologies. However, till today, the relation of activity distribution to attenuation remains a challenging frontier, and the occurrence of unavoidable artifacts persists (28).

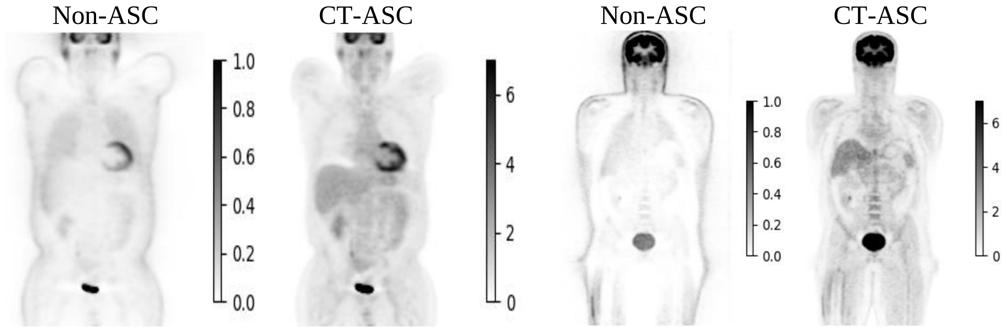


Figure 1: Two examples of PET images before and after ASC. The non-ASC shows PET scans without attenuation and scatter correction; the CT-ASC shows PET scans after applying CT-based attenuation and scatter correction. As can be seen, the corrected images have higher intensity values due to the attenuation correction procedure that compensates for part of a signal, which was lost during photons passage through the body. So, more accurate representations of the radioactive tracer distribution are shown, and more explicit images contain information about anatomic structures and lesions. Notice a multiplied clarity and enhanced lesion detection in the CT-ASC images compared to the non-ASC images.

Recent progress in segmentation, classification, detection, noise reduction, and reconstruction research areas utilizing artificial intelligence (AI) has encouraged nuclear medicine researchers to investigate CT-free methods for ASC in PET imaging (49–68). CT Elimination has advantages for those who require repeated scans, especially for pediatric patients, as even a minor reduction in cumulative radiation exposure could cause a significant impact (69,70).

Some deep learning-based methods have been developed to generate the synthesis of pseudo-CT images from MRI or uncorrected PET data, prediction of scatter maps from emission data (71–76), while other research focuses on the direct generation of ASC PET images from non-attenuation-scatter-correction (NAC) as inputs to predict ASC PET images directly (50,54,77). The direct image-to-image translation technique not only highlights the capabilities of DL models in ASC without CT but also possesses the ability to detect and correct artifacts in PET images accurately (78,79).

An important question facing researchers today is the practical applicability of these models in clinical environments. Due to differences in spatial resolution, sensitivity, and technical information among scanners and variations of radiotracer biodistribution in the body, a model optimized for data from one specific scanner may not perform effectively under different conditions or other equipment. Moreover, not all medical centers are equipped with a dedicated AI team or even restricted in data sharing by ethical and regulatory considerations (80–83). Federated learning (FL) addresses some challenges, such as data privacy and limited dataset sizes in medical imaging (27,28,84,85). Yet, to achieve widespread clinical acceptance and enhance PET imaging's diagnostic capabilities, novel correction techniques in CT-free PET imaging avenues must be sought.

Previous research has shown that direct ASC frameworks can correct artifacts in ^{18}F-FDG PET/CT images (76). A DL model using the idea of decomposition of PET image into two anatomical independent and dependent information was proposed to address the limitations across large heterogeneity of tracers and scanners of PET imaging (49). Additionally, the detection and correction of ^{18}Ga image artifacts using a tuned direct ASC model for multiple centers have been assessed (27). Despite these advances, further investigation into a multi-center model for quantitative analysis of gallium studies is still needed.

The main aim of this study is to predict CT-based attenuation-scatter correction (MAC) of PET images based only on the non-attenuation-scatter correction (NAC) image using DL. So, we addressed the direct ASC of PET images without using anatomical information from CT and evaluated the performance of the model in removing image artifacts in a multi-center dataset. As part of our objective, we also tested the capability of the proposed idea, the anatomical independent and dependent information (49) to create a universal model. We estimated and compared the performance of models in different radiotracers and scanners.

Material and methods

Datasets

^{68}Ga PET/CT scans from five different hospitals from a previous study (27) were used for training and initial model validation in the primary stage of our study. A secondary dataset (28), distinct in both the imaging centers and the type of radiotracer used (^{18}F-FDG PET scans from two different hospitals), was incorporated to test the model's adaptability. Additionally, a specialized set of images presenting artifacts (27) was included to assess the model's capability to detect and correct image quality issues.

^{68}Ga PET/CT dataset

A cohort of 1000 patients underwent ^{68}Ga-prostate-specific membrane antigen (PSMA) PET/CT imaging across five centers located in different countries (27). To ensure the integrity of the data for model training, an expert in nuclear medicine evaluated all the scans, identifying 184 images of optimal quality without artifacts from the total pool (27). Detailed information on the datasets collected is outlined in Table 1. The CT-based ASC was applied to amend PET images for accurate correction of attenuation and scatter effects on the images. For this study, non-attenuation-corrected images will be referred to as NAC, and CT-based attenuation scatter-corrected images will be denoted as MAC.

Table 1: Data information in 5 different imaging centers.

Center	No	Train	Validation	Test	Scanner	Reconstruction	Matrix size × Z*
Center 1	56	43	11	2	Siemens Biograph 6	3D-OSEM	168 × 168
Center 2	31	25	4	2	GE Discovery IQ	3D-OSEM	192 × 192
Center 3	45	35	8	2	Siemens mCT	3D-OSEM	200 × 200
Center 4	40	28	10	2	Siemens Biograph 6	3D-OSEM	168 × 168
External Center	12	-	-	12	Siemens Horizon	PSF+TOF+3D-OSEM	180 × 180
Total	184	131	33	20	-	-	-

* Z' representing the number of slices in the axial view, depends on body length, scanner resolution, scan protocol, and patient positioning. Therefore, it varies for each patient.

^{18}F-FDG Datasets

To assess the model's performance with different radiotracers, our study incorporated a dataset of 98 whole-body ^{18}F-FDG PET scans originating from two distinct centers, representing our external radiotracer dataset (Figure 2) (28). During the preprocessing phase, the intensities of voxels in both MAC and NAC images were standardized for SUVs by scaling factors, 9 for MAC and 3 for NAC images.

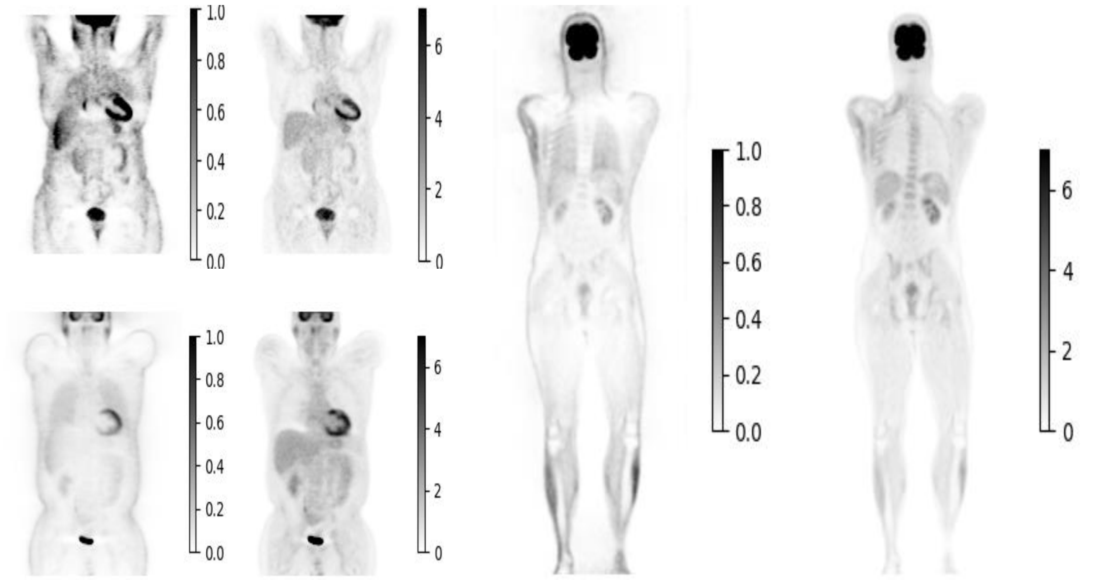


Figure 2: Sample of coronal slices from an FDG dataset, illustrating the range in axial slice counts, which vary from 180 to 600 based on the organ of interest.

Table 2: Overview of External Radiotracer Dataset Specifications.

Center	No	Train	Validation	Test	Matrix size × Z
Center 6	55	39	6	11	272 × 200
Center 7	43	23	9	10	272 × 200
Total	98	62	15	21	-

* Z' representing the number of slices in the axial view, depends on body length, scanner resolution, scan protocol, and patient positioning. So, it is different for each patient.

Artifact dataset

A third test set was utilized to evaluate the performance of the developed model under more challenging conditions. This set consisted of imaging data from 198 patients, each displaying various types of artifacts. The artifacts in this dataset were chosen to test how well the model can handle and correctly interpret images that are distorted by common problems seen in clinical ^{18}Ga imaging, like motion and Halo artifacts.

Data Preparation

Normalization of ^{68}Ga PET Image

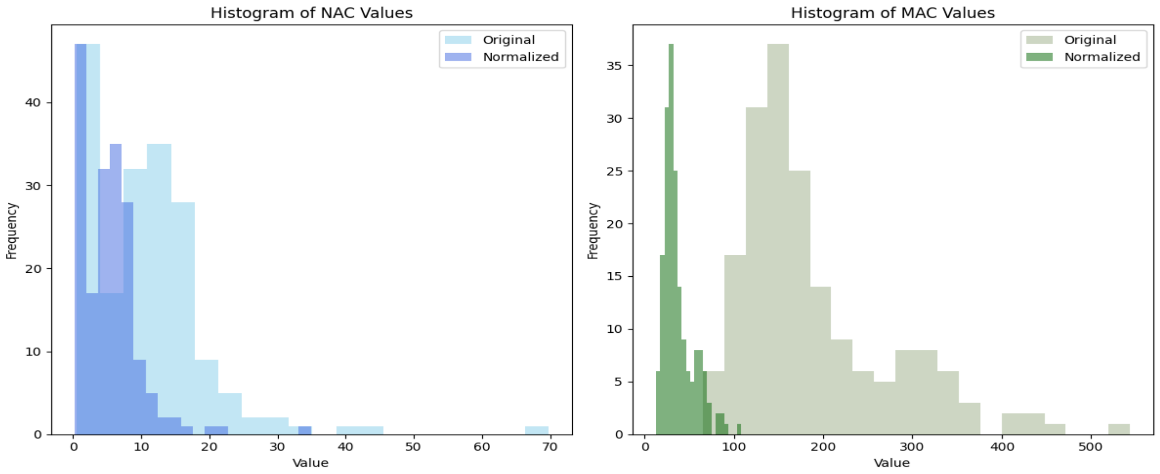
The standard uptake value (SUV) in PET imaging is an important standardization procedure that allows quantitative measurement. This means that the detected radiotracer concentration reflects the metabolism of the patient's body. It corrects based on the radiotracer injected dose and the patient's body weight. This conversion is essential as it factors in variations due to patient size and the amount of radiotracer administered. The SUV is calculated using the Equation 1:

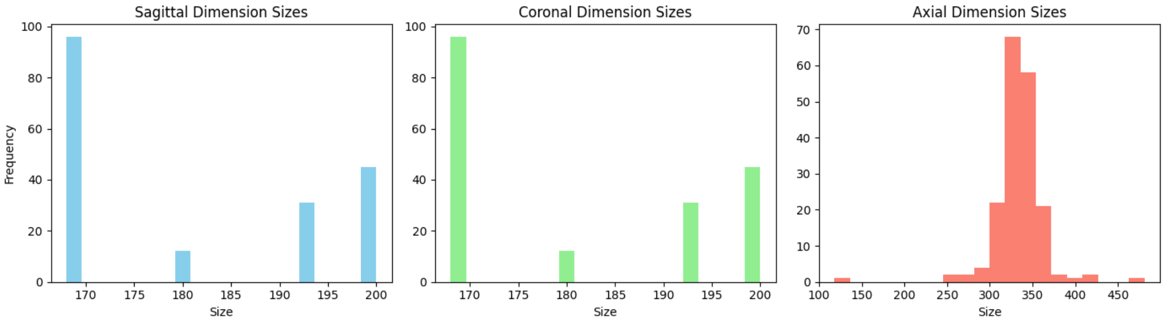
$$SUV = \frac{Voxel\ Activity\ Concentration_{(Bq/ml)}}{Injected\ Dose_{(Bq)} \big/ Body\ Weight_{(kg)}} \qquad (1)$$

To turn the voxel values into SUV metrics, this conversion was done the same way on all MAC and NAC images. To preserve quantitative values across all images and since DL models operate more efficiently with smaller numbers, the images were normalized by dividing them by a constant factor. For the ^{68}Ga dataset MAC images underwent a factor of 5 scaling, while 2 was picked for NAC images based on previous studies (27). The histogram of the images post-normalization illustrates the effect of this scaling on the distribution of voxel intensities, confirming the consistency of intensity levels across the processed images (Figure 2A).

Data Transformation and Augmentations:

For training data preparation, each PET image was initially, followed by the addition of zero-padding to standardize the dimensions to a uniform bounding box size of 168×168×Z (with 'Z' representing the count of slices), as illustrated in Figure 3a. This ensured the retention of the original image resolution and anatomical structure. To ensure uniformity and enhance the training process's efficiency, all ^{68}Ga PET images were re-scaled to a voxel size of 4.07×4.07×3.0 mm^3, the most common resolution across the collected data and crucial for consistent image analysis. This standardization was crucial for achieving consistent image quality throughout the dataset. Details regarding the initial voxel spacing are provided in Figure 3b.





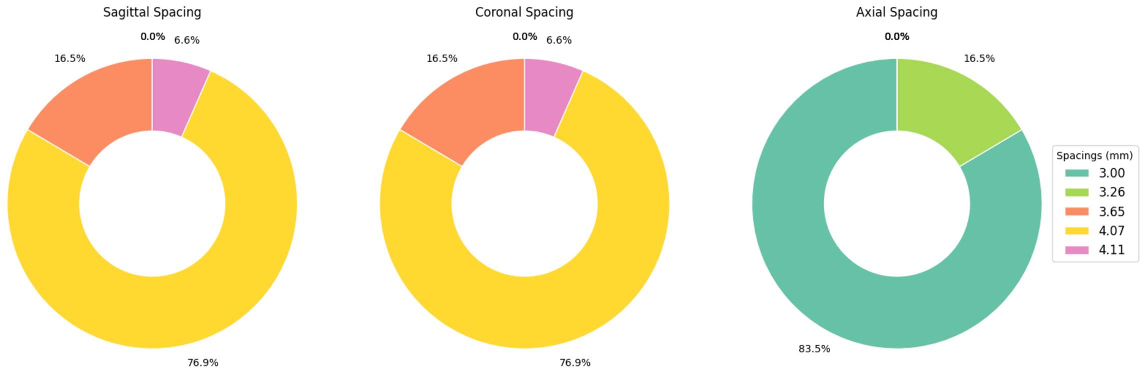


*Figure 3: **A)** Distribution of maximum intensity values for NAC and MAC images, displaying variations pre- and post-normalization to highlight data scaling effects. NAC images were scaled down by a factor of 2 and MAC images by a factor of 5. **B)** initial PET image dimensions are distributed across sagittal, coronal, and axial planes. Each bar represents the frequency of occurrence for specific dimension sizes within the dataset. **C)** Proportion of different voxel spacings utilized in PET image preprocessing. The donut charts depict the percentage of images corresponding to each voxel spacing dimension in millimeters across sagittal, coronal, and axial views.*

Generation of Anatomy-Dependent Correction Maps (ADCM)

In exploring advanced techniques for PET image correction, we examine a decomposition-based deep learning approach previously proposed (49). Based on this idea, the attenuation scatter corrected image could be divided into two parts: an anatomy-independent part (which contains information related to tracers' distribution and diseases) and the anatomy-dependent part (which contains anatomical information about the body). In other words, in this method MAC image has been formed from these two key components. Anatomy-independent information, which correlates with tracer type and disease pathology, and anatomy-dependent map necessary for ASC, named ADCM. If the model is trained on ADCM, we simply could achieve the DL-MAC by multiplying the DL-ADCM by NAC images

So, the ADCM for each voxel is defined by conditional Equation 2, which captures the ratio of the MAC intensity to the NACs:

$$\begin{gathered} If\ PET_{NAC}[x,y,z] \geq \varepsilon\ then \\ PET_{ADCM}[x,y,z] = {PET_{NAC}[x,y,z]}\big/{PET_{MAC}[x,y,z]} \\ else \quad PET_{ADCM}[x,y,z] = PET_{MAC}[x,y,z] \end{gathered} \qquad (2)$$

The threshold ε (ε = 0.001) ensures that division by zero is avoided, defaulting to the MAC intensity where necessary.

In the evaluation phase, our trained model predicts the DL-ADCM for a given NAC. We then employ the following transformation (Equation 3) to achieve the DL model-based attenuation correction (DL):

$$\begin{gathered} If\ PET_{NAC}[x,y,z] > \varepsilon\ then \\ PET_{DL}[x,y,z] = PET_{NAC}[x,y,z] * PET_{DL-ADCM}[x,y,z] \\ else\ \ PET_{DL}[x,y,z] = PET_{NAC}[x,y,z] \end{gathered} \qquad (3)$$

Sample cases are visualized in Figure 4.

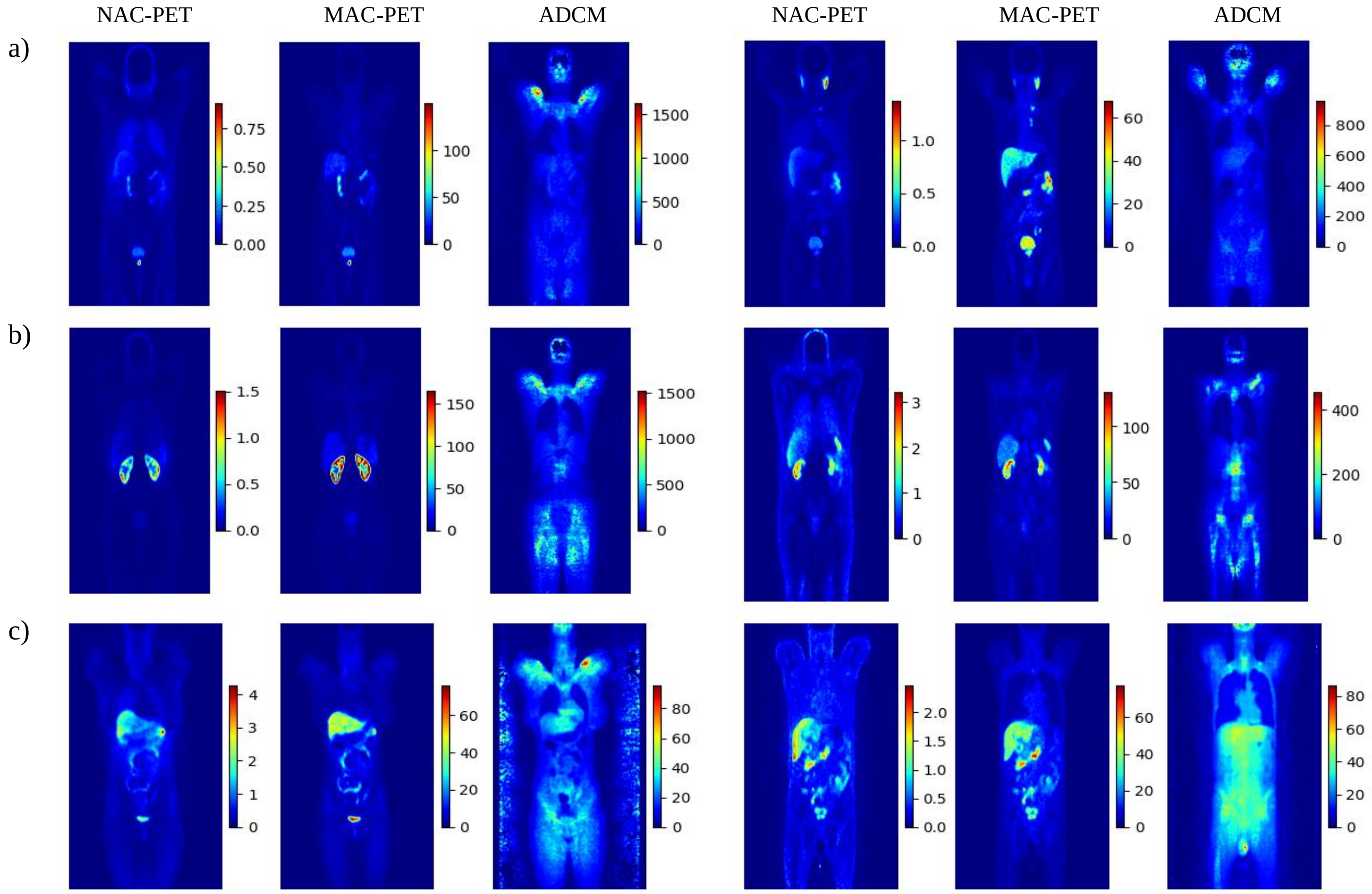


Figure 4: The middle slice of the coronal view for NAC, MAC, and ADCM images. Color bar unit: SUV

Normalization of ADCM

As we already mentioned, to calibrate ADCM to preserve the quantitative accuracy of SUV, which is necessary for accurate clinical interpretations. We came up with an empirical normalization factor of 50 for ADCM values.

.Deep neural network

We used the Dyn-UNet architecture, well-known for its adaptability and efficiency in processing biomedical images (86). This model is chosen for its dynamic configuration and deep supervision. The Dyn-UNet model's initialization is designed to find the best kernel sizes and strides based on the size and spacing of the input patches in our dataset. These parameters were set by evaluating the spatial dimensions and resolution of the input data, ensuring the network architecture is perfectly aligned with the inherent characteristics of our medical images.

For the ^{68}Ga dataset, the computed kernel sizes and strides are set to four layers of [3, 3, 3] kernels, with strides transitioning from [1, 1, 1] in the initial layer to [2, 2, 1] in the deeper layers, based on initializer's suggestion. The model has a deep architecture with 124 layers of convolutional, instance normalization, and activation layers, indicating a typical UNet-like structure with down-sampling and up-sampling paths, as shown in Figure 5. The channels or width of the model varies from 32 to 256 in deeper layers, with 10,934,373 as the total parameters. Additionally, the implementation of deep supervision, with two supervision heads[1] (87), enhanced the learning process by optimizing the network's final and intermediate layers. By setting the ReLU activation function in the last layer, we get the non-zero value for the concept of the PET image. Our deep learning network was designed to process NAC images as inputs to generate MAC or ADCM images for different approaches and will be elaborated upon later. We also used L1 regularization in the training process to prevent overfitting. L1 regularization introduces a penalty proportional to the absolute value of the model parameters, hence favoring sparsity in the weights. This will smooth the complexity of the model and help prevent generalization by avoiding overfitting. We used an L1 regularization strength of 0.0001.

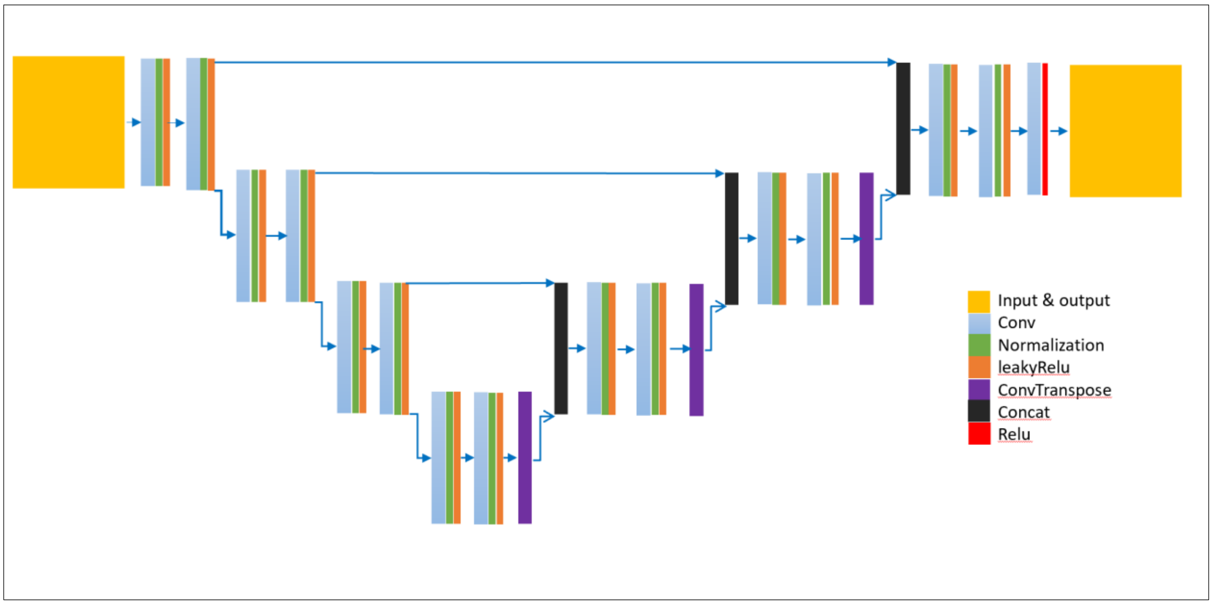


Figure 6: The architecture of DynUnet.

Network training involved using 3D patches sized at 168×168×16 and 20 sample patches per patient. The key training parameters were as follows: Learning rate of 0.001, Loss function of the mean squared error (MSE)—also referred to as the squared L2 norm. The MSE loss function was employed to measure the deviation of the network's output from the MAC ground truth.

The network was optimized using the Adam optimizer. The beta coefficients, set at 0.5 and 0.999, controlled the moment estimates' exponential decay rates. Only artifact-free datasets were used during the network's training and validation stages to maintain the model's integrity. We trained the network for 500 epochs with a batch size of 4 to ensure adequate convergence and comprehensive learning from the dataset. To prevent data leakage and ensure data integrity, patients were not overlapped across the training, testing, and validation datasets, maintaining the independence of each dataset. Details on alternative models tested, including those that did not meet our criteria for inclusion in the final report, are documented in Supplementary Material 1 for transparency and completeness.

[1] Deep_supervision: in training mode, make the forward function output not only the final feature map, but also from the intermediate up sample layers. So, all intermediate feature maps are interpolated into the same size as the final feature map and stacked together as one single tensor (87).

Training approaches for deep learning models:

Integrated multi-center model (IMCM):

A Dyn-Unet DL model was developed using a combined dataset from four different centers, all utilizing ^{68}Ga-based radiotracers. This model was initially trained on a collective dataset and subsequently tested on an external center's data to evaluate its generalization capabilities. It was also tested within the originating dataset from each center. This approach aims to overcome the limitations of models trained on data from single centers, which may struggle with generalizability to new, unseen cases. The training and validation losses for the IMCM are illustrated in Figure 6. The model was trained and validated on NAC, and the ground truth was MAC.

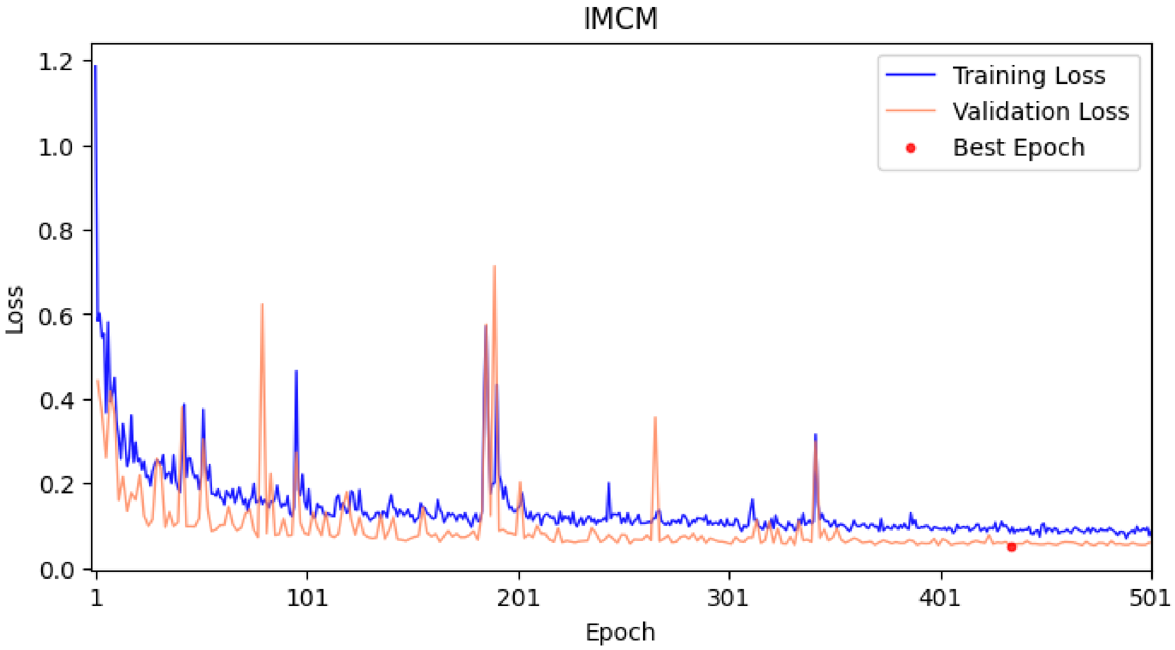


Figure 7: Training and validation loss for the Integrated Multi-Center Model showing the best metric of 0.0527 at epoch 434.

Anatomy-Dependent Correction Model (ADCM):

This methodology adopts a new approach from NAC to MAC with ADCM in between as ground truth. So, the model was trained only just on anatomy-dependent components of data, or ADCM (Equation 2). And using Equation 3 the DL version of MAC images was generated.

The previous DynUNet network was used to evaluate this approach. This model's effectiveness is evaluated through its ability to generalize across different centers and tracers, testing its robustness in a variety of clinical settings. The training progress and validation stability for the ADCM are detailed in Figure 7.

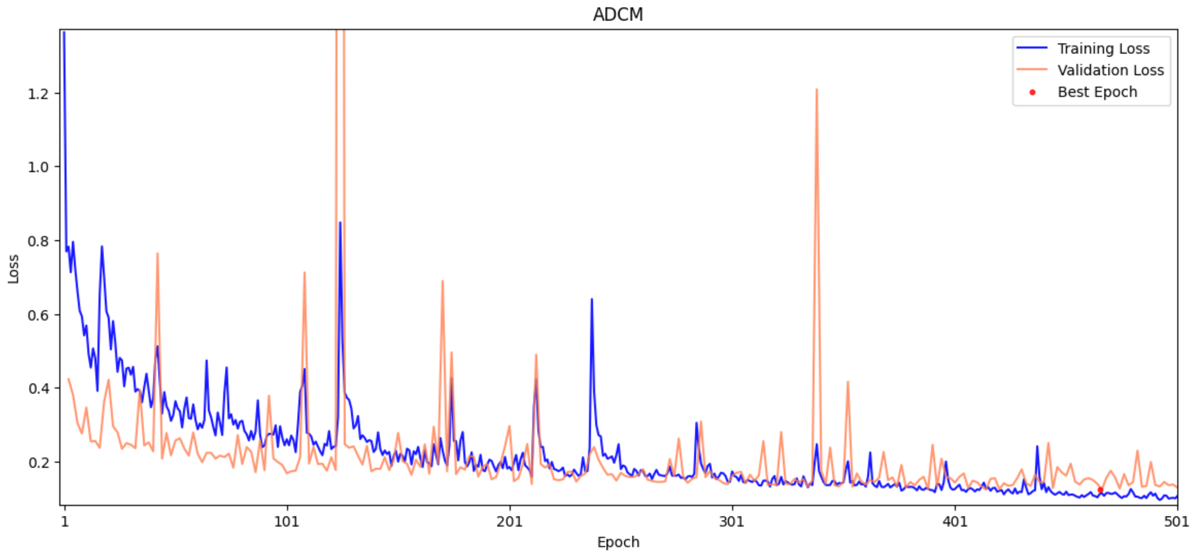


Figure 8: Training and validation loss for the ADCM model, where the best metric of 0.1237 was reached at epoch 466.

Tuned Transfer Learning for IMCM model (TL-MC):

The IMCM model underwent tuning through transfer learning (TL) to address the challenges encountered with different radiotracers. This method involves modifying the DL model by integrating learning with the new dataset. By retaining all weights and choosing a very low learning rate, we allowed the network to adapt itself, called fine-tuning the whole model. It is assumed that the data distribution has changed, and model needs to adapt features for this new distribution. This refinement enhanced the model's performance and adaptability across different tracer types, providing a more robust solution that could potentially handle variability more effectively. The effectiveness of the TL approach is depicted in Figure 8, demonstrating rapid convergence and effective transfer learning.

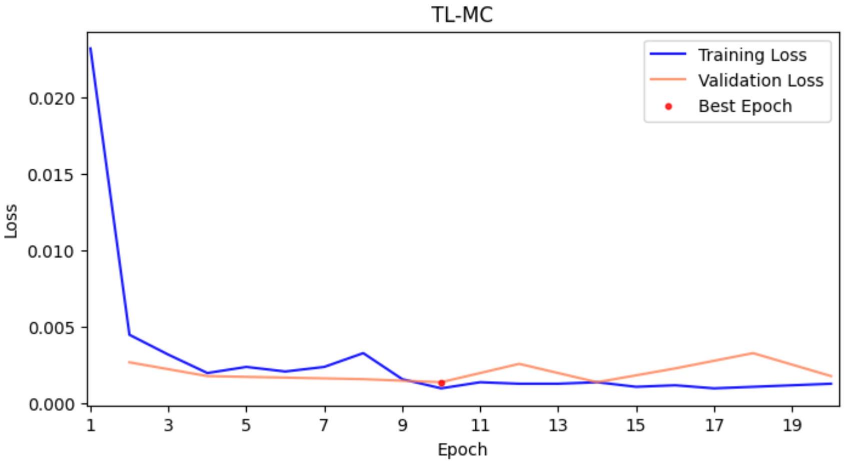


Figure 9: Training and validation loss for the Tune TL Model with a best metric of 0.0014 achieved at epoch 10, demonstrating rapid convergence and effective transfer learning.

Quantitative evaluation:

The model's efficacy was rigorously quantified using a range of statistical metrics, calculated by comparing the DL-predicted PET images against the ground truth CT-based attenuation/scatter-corrected images. These voxel-wise metrics are computed as follows:

- **Mean Error (ME):** This reflects the average deviation across all voxels.

$$ME = \frac{1}{tot}\sum_{v=1}^{tot} PET_{pred}(v) - PET_{ref}(v) \qquad (4)$$

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors without considering their direction.

$$MAE = \frac{1}{tot}\sum_{v=1}^{tot}\left|PET_{pred}(v) - PET_{ref}(v)\right| \qquad (5)$$

- **Relative Error (RE%):** Provides a percentage error relative to the true values, indicating the proportion of the deviation.

$$RE\ (\%) = \frac{1}{tot}\sum_{v=1}^{tot}\frac{PET_{pred}(v) - PET_{ref}(v)}{PET_{ref}(v)} \times 100\% \qquad (6)$$

- **Root Mean Squared Error (RMSE):** Measures the average of the squared differences between the predicted and reference values. It is useful for quantifying the deviation in predictions from the observed values across the dataset.

$$RMSE = \sqrt{\frac{1}{tot}\sum_{v=1}^{tot}(PET_{pred}(v) - PET_{ref}(v))^2} \qquad (7)$$

This refers to the total number of voxels and PET_{pred} and PET_{ref} indicate the predicted image via DL model and the ground truth image, respectively.

- **Peak Signal-to-Noise Ratio (PSNR):** Evaluates the ratio of the maximum possible signal to the corrupting noise.

$$\mathrm{PSNR(dB)} = 10\log_{10}(\frac{\mathrm{Peak}^2}{\mathrm{MSE}}) \qquad (8)$$

In Eq. 8, Peak represents the maximum intensity value in the image.

- **Structural Similarity Index (SSIM)** (88)**:** Assesses the perceptual quality of the predicted images relative to the reference images.

$$SSIM\left(PET_{pred}, PET_{ref}\right) = \frac{(2\mu_{pred}\mu_{ref} + c_1)(2\sigma_{pred,ref} + c_2)}{\left(\mu_{pred}^2 + \mu_{ref}^2 + c_1\right)(\sigma_{pred}^2 + \sigma_{ref}^2 + c_2)} \qquad (9)$$

where: μ_{pred} and μ_{ref} are the averages of the pixel intensities in the predicted PET images (PET_{pred}) and the CT-attenuation corrected PET images (PET_{ref}), respectively. σ_{pred}^2 and σ_{ref}^2 are the variances of the pixel intensities in the predicted and CT-attenuation corrected PET images, respectively. $\sigma_{pred,ref}$ is the covariance of the predicted and CT-attenuation corrected PET images. $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are constants to stabilize the division with a weak denominator; L is the dynamic range of the pixel values (typically $2^{bit\ per\ pixel} - 1$). k_1= 0.01 and k_2=0.03 is the default value for the stabilization constants.

Results

Quantitative assessment

Cross-Center Results:

This section evaluated the two proposed DL algorithms on the ^{68}Ga-PET dataset (IMCM and ADCM). We tested the trained DL model with two internal and external test sets to evaluate its robustness. The internal test sets included 8 subjects from 4 different centers as an external test set and 12 subjects from an external, non-seen center. Figure 9 displays the quantitative accuracy of the DL-based images compared to the ground-truth MAC images for internal and external centers. The visualization of DL images to the expert demonstrate that both methods effectively performed some degree of attenuation and scattering correction across these centers. The statistical assessment showed that IMCM outperformed ADCM (p-value<0.02). Refer to the Supplementary Material in Figure 1 for a detailed center-wise analysis.

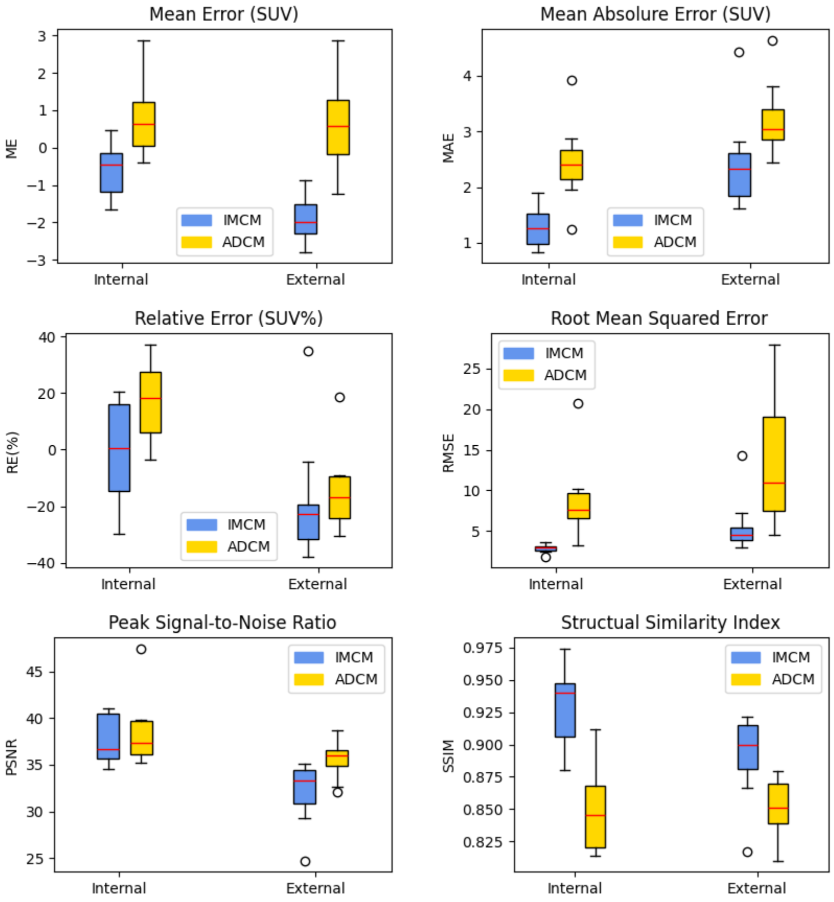


Figure 10: Quantitative metrics for the IMCM and ADCM methods across internal and external centers, including mean error (SUV), mean absolute error (SUV), relative error (SUV%), root mean squared error, peak signal-to-noise ratio, and structural similarity index.

For the external center, ADCM yielded an ME of -0.63±0.96 and an MAE of 3.072±1.01. In contrast, the IMCM demonstrated improved consistency with an ME of -1.83±1.39 and an MAE of 2.59±0.93. Internal centers showed that ADCM produced an ME of 0.37±1.45 and an MAE of 2.34±0.77, while IMCM shows lower error in both metrics. PSNR favored the ADCM method, with 35.53±2.12 compared to 38.25±1.92 for the IMCM method. Notably, SSIM was superior for IMCM in the external center, at 0.88±0.020. Details are available in Supplementary Material 2, Table 7.

In addition to voxel-wise assessments, model performance was further validated through various statistical tests, which compared image-derived metrics between different training models. The Wilcoxon test was used due to the data's non-normal distribution, as evidenced by the Shapiro-Wilk tests. The Wilcoxon test showed that the ADCM and IMCM datasets were significantly different for all metrics except RE (SUV%), where the p-value does not indicate a statistically significant difference with a p-value higher than 0.05. IMCM consistently shows lower errors, a higher PSNR, and higher SSIM values, indicating superior image quality and more reliable estimations. These findings are further detailed in Supplementary Material 2, Table 8 and 9.

In the joint histogram analysis, Figure 10 visualizes the voxel-wise correlation across the different centers for both methods. A clear difference in predictive accuracy and linearity in SUV estimation was demonstrated. In the external center, the IMCM regression slope of 0.65 ± 0.02 with an R-value of 0.949 clearly showed a systematic underestimation over the range of predicted SUV values, compared to ADCM, which showed a slope of 1.18 ± 0.10 and an R^2 of 0.850.

In internal centers, the behavior of the methods differed, with the IMCM method being closer to the ideal prediction, especially evident at center C3 with a regression slope of 0.87 ± 0.01 and an R^2 of 0.988. On the other hand, the ADCM method had slopes greater than one in some cases (1.13 ± 0.03 at C2 and 1.19 ± 0.03 at C4).

The voxel-wise analysis further confirmed these findings, showing larger discrepancies in centers where ADCM predicted significantly higher values. Overall, these results demonstrate that ADCM appears to be closer to the truth in some centers because the R^2-values are higher. However, the reliability and clinical usefulness of ADCM can be called into question. IMCM demonstrated image quality comparable to MAC and preserved more detailed information with lower noise compared to ADCM.

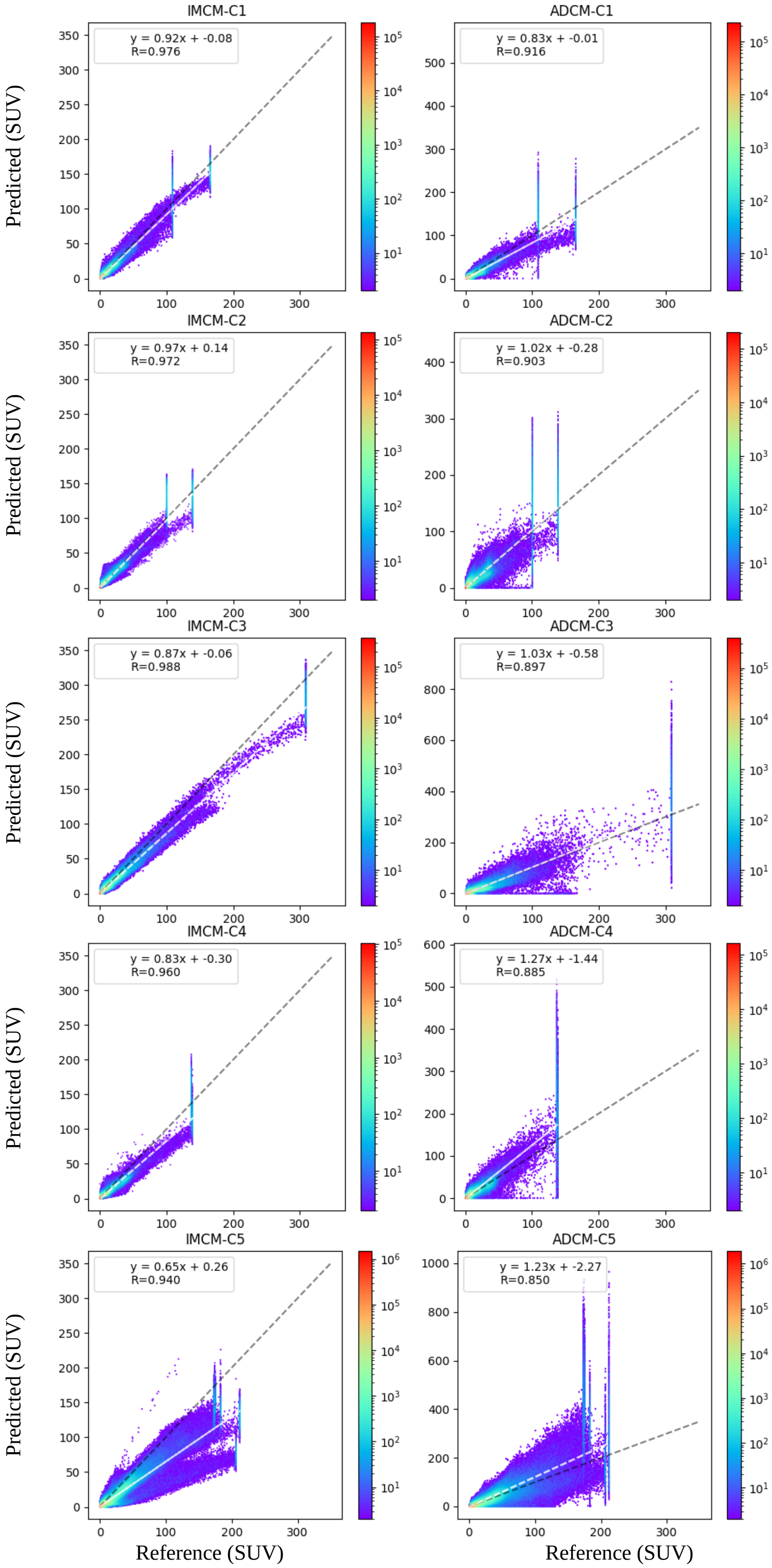


Figure 11: Joint histogram analysis displaying the correlation between activity concentration in DL-IMCM and DL-ADCM images versus reference MAC images serving as the ground truth. Note that a logarithmic color scale was used to display the number of voxels. C1-4 are internal centers, while C5 is an external center.

Cross-Tracer Results:

As part of our assessment of generalization capabilities across different tracer types, IMCM was initially tested without specific tuning for cross-tracer variations. As proved before, the results revealed that the IMCM, without prior tuning, struggled to maintain its efficacy when applied to different radiopharmaceutical tracers (76)(52). As a part of the generalization assessment among different tracer forms, IMCM was first tested without specific tuning for cross-tracer variation. As mentioned before, the results show the IMCM model, without prior tuning, struggled to maintain its performance when applied to different radiopharmaceutical tracers (39).

So, the ^{18}F-FDG-PET dataset was used as a cross-tracer in this study to test the two proposed DL algorithms: TL-MC (the tuned version of IMCM) and ADCM. We tested the trained DL model to evaluate its robustness, which included 20 subjects from 2 different centers as external centers. Figure 11 showcases a sample coronal slice of IMCM, TL-MC, and ADCM on cross-tracer subjects. The significant drop in accuracy and increased error rates highlight the challenges in achieving robust cross-tracer generalization with a single, unified model approach.

The two approaches, TL-MC and ADCM, indicate significant differences in error metrics. Both ME and MAE indicated much smaller error margins for the TL-MC, with the overall mean values reflecting better accuracy than the ADCM. The TL-MC ME deviated narrowly by -0.10±0.76, while the ADCM deviated by 0.82±0.70, signifying a much wider spread of the SUV estimates (Figure 12). These are shown as RE%. This also confirms that TL-MC had a better performance. The RE spread was relatively lower for TL-MC, averaging at 30±50%, in contrast with ADCM, where the spread was much broader at 50±100%.

TL-MC gave a lower RMSE of 2.0 ± 0.6, which pointed out consistency and reliability compared to ADCM's 3.2 ± 1.1. It was also better than ADCM in image quality metrics, with higher PSNR and SSIM values, showing tighter control over noise and structural fidelity.

Altogether, these findings point towards superiority in the use of TL-MC over ADCM in terms of accuracy and consistency in all major key PET imaging metrics, and the use of this approach is recommended in clinical practice where precision is critical. The data makes a compelling case that TL-MC should be preferred with respect to its strong performance in consistently keeping lower errors in the images. For a comprehensive view and deeper analysis, refer to the box plots in Figure 12. Detailed statistical comparisons of these metrics are illustrated in Supplementary Material 2, Tables 10 and 11, provided.

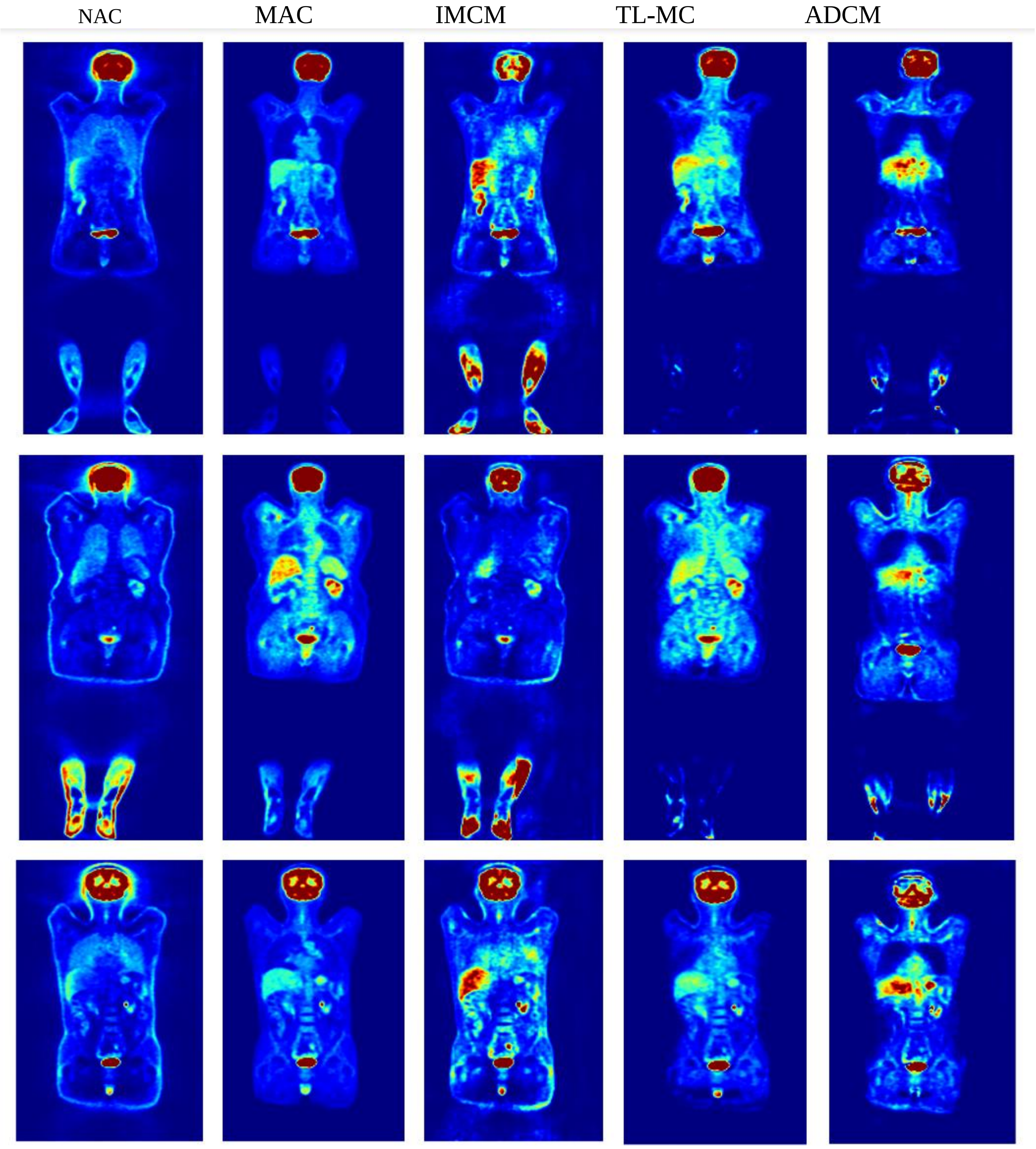


Figure 12: From left to right, a coronal slice of NAC, MAC, IMCM, TL-MC, and ADCM on cross-tracer subjects, respectively.

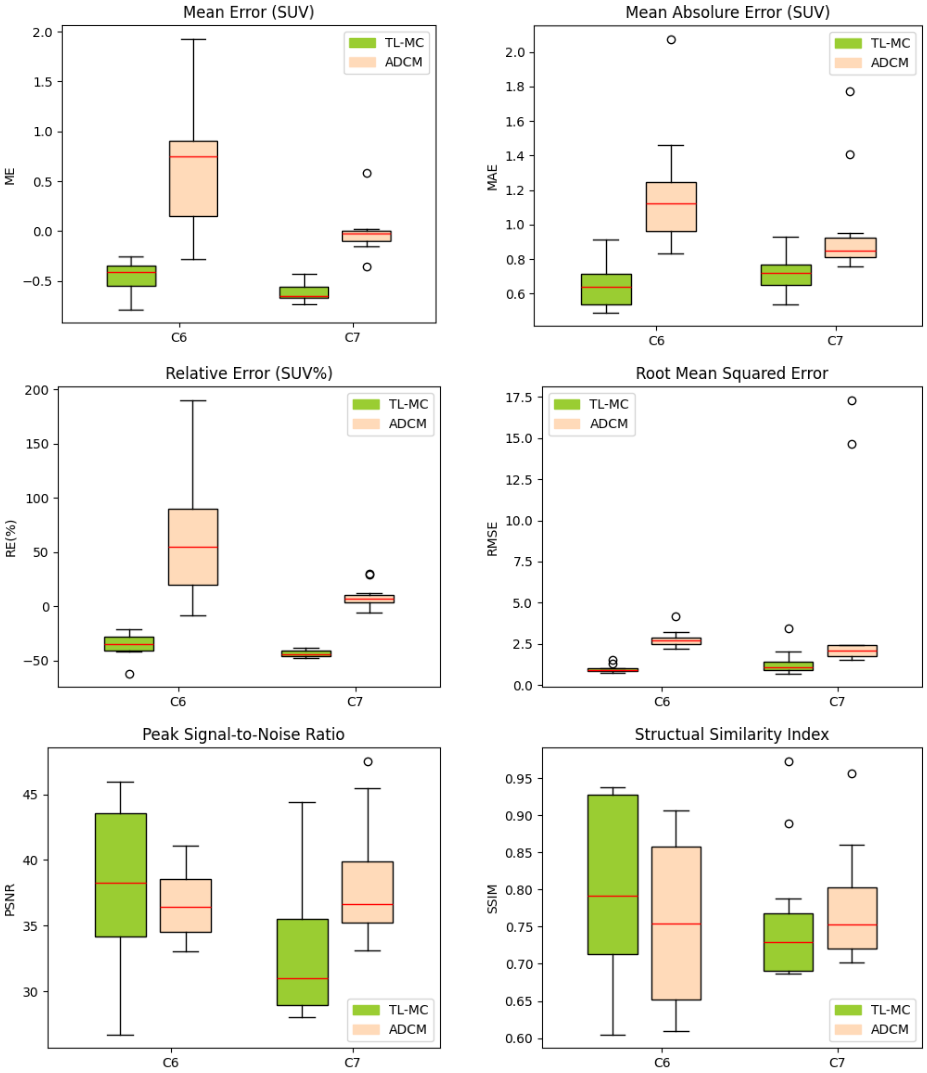


Figure 13: Comparative Analysis of Imaging Metrics Between ADCM and IMCM Methods. The box plots depict the distribution of mean error (SUV), mean absolute error (SUV), relative error (SUV%), root mean squared error, peak signal-to-noise ratio, and structural similarity index across centers C6 and C7.

Further investigation through joint histogram analysis of the TL-MC and ADCM models in different centers provides a precise understanding of each model's predictive capabilities for SUVs. The TL-MC model closely matches reference values, as evidenced by regression slopes of 0.98 ± 0.38 and 0.69 ± 0.08 at two respective centers. Notably, this model also has high correlation coefficients of 0.915 and 0.918, emphasizing its precision in SUV prediction despite a tendency to slightly underestimate values, particularly at Center C7, as presented in the analysis.

On the other hand, the ADCM model has lower correlation coefficients of 0.660 and 0.678, even though its regression slopes are higher at 1.10 ± 0.46 and 1.35 ± 0.66, which means it overestimates the data. This discrepancy highlights the lesser consistency and reliability of its predictions when compared to TL-MC. Contrary to intuitive expectations of better correlation, the higher slopes observed in ADCM indicate a greater deviation from the reference line, pointing to a systematic error in overestimating SUVs.

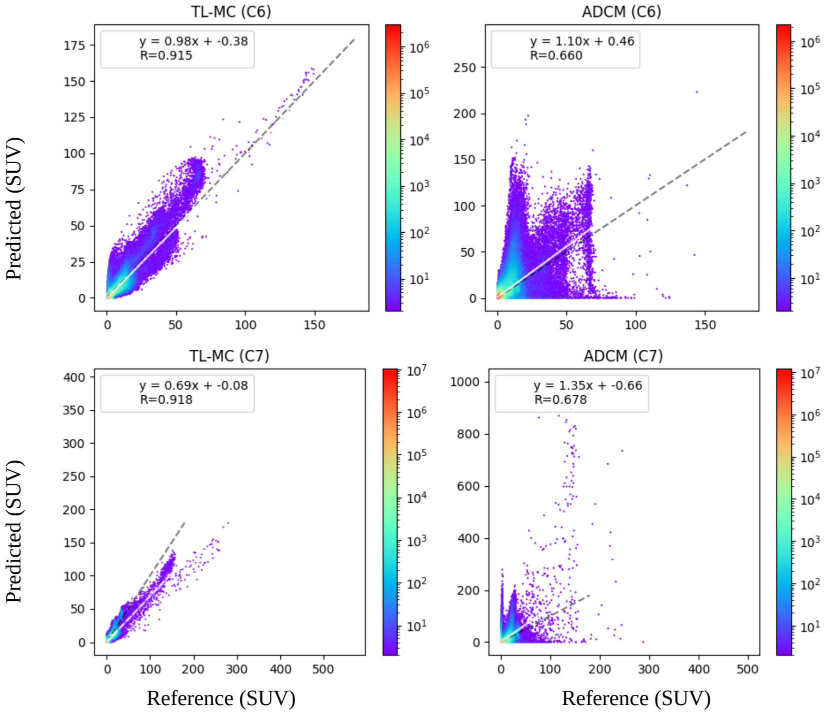


Figure 14: Joint histogram analysis displaying the correlation between activity concentration in TL-MC and ADCM images versus reference MAC images serving as the ground truth for cross-tracer. Note that a logarithmic color-scale was used to display the number of voxels.

Case Study on Artifact Images

This section examined a series of case studies involving repeated scans. These repeated scans have been requested by nuclear medicine physicians shortly after initial assessments. Figures 14, 15, and 16 display the imaging results for patients with halo artifacts in the pelvic, kidney, diaphragm, lung, liver, and spleen regions. These artifacts were removed in the repeated scan. The ICMC method produced artifact-free images of high quality, diagnostic confidence, and nearly identical to the initial scan. Figure 17 features patients with a halo artifact in the kidneys. A repeated scan was conducted in this region due to the initial scan's low image quality and diagnostic confidence. Unfortunately, in some cases, the repeated scan could not remove these artifacts. Nonetheless, the ICMC model successfully eliminated the artifact in both the original and subsequent scans.

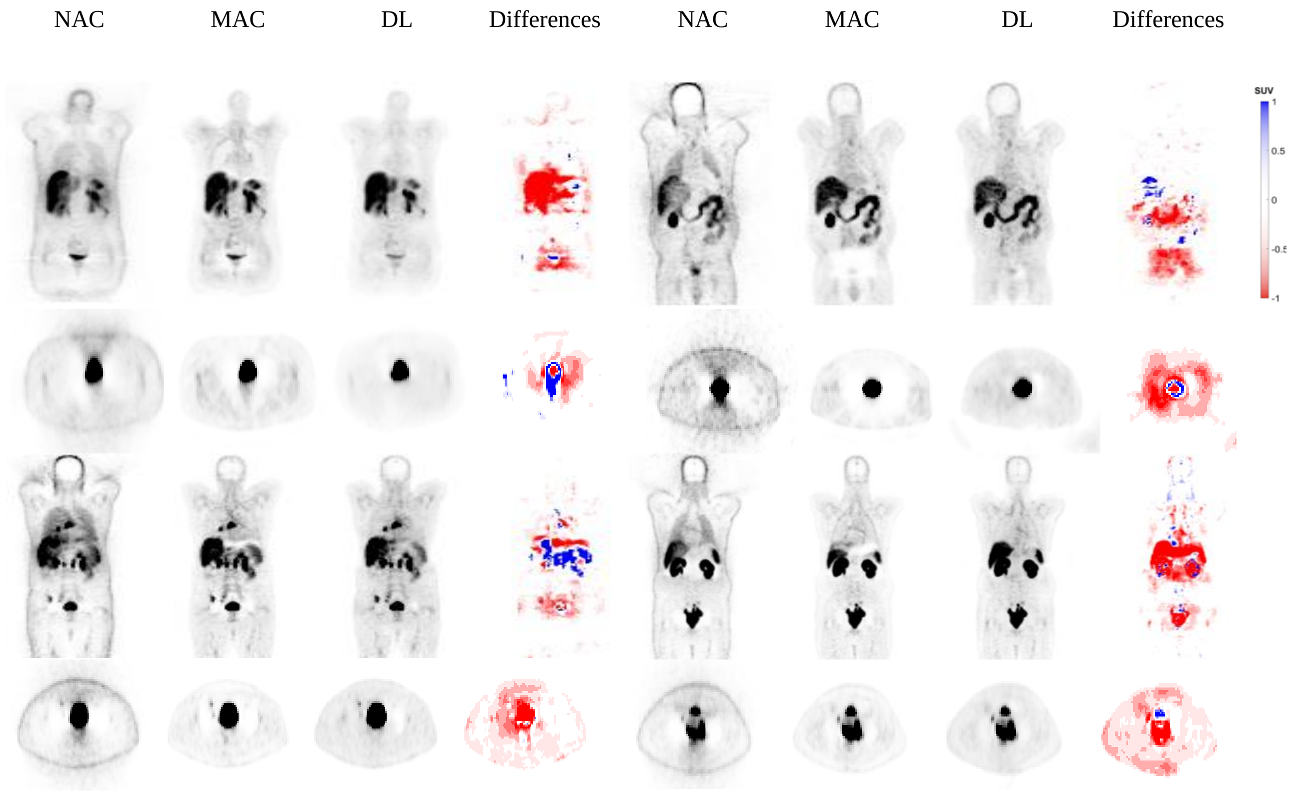


Figure 15: Coronal and axial views of 4 clinical studies showing from left to right NAC, MAC, IMCM-DL and the difference images of MAC and DL image. The images generated using the IMCM approach successfully corrected the halo artefact in pelvic area. The red intensity regions represent areas where the DL image contains more information compared to the MAC. These regions are artifactual parts of the MAC images.

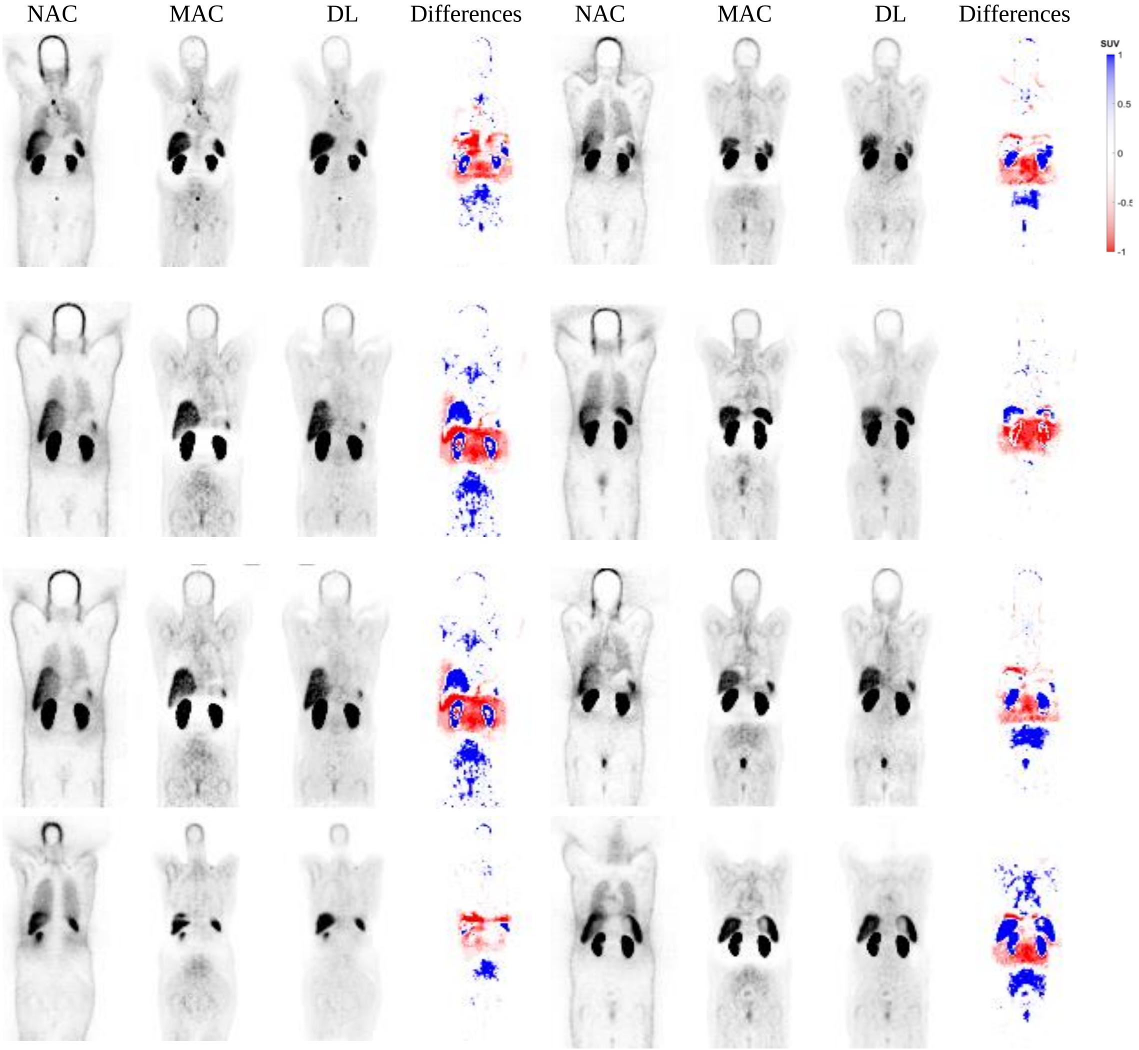


Figure 16: Coronal views of 8 clinical studies, representing from left to right: NAC, MAC, IMCM-DL and the difference images of MAC and DL image. Our method effectively disentangles halo artefacts in the kidney area. The red intensity regions represent areas where the DL image contains more information compared to the MAC. These regions are artifactual parts of the MAC images.

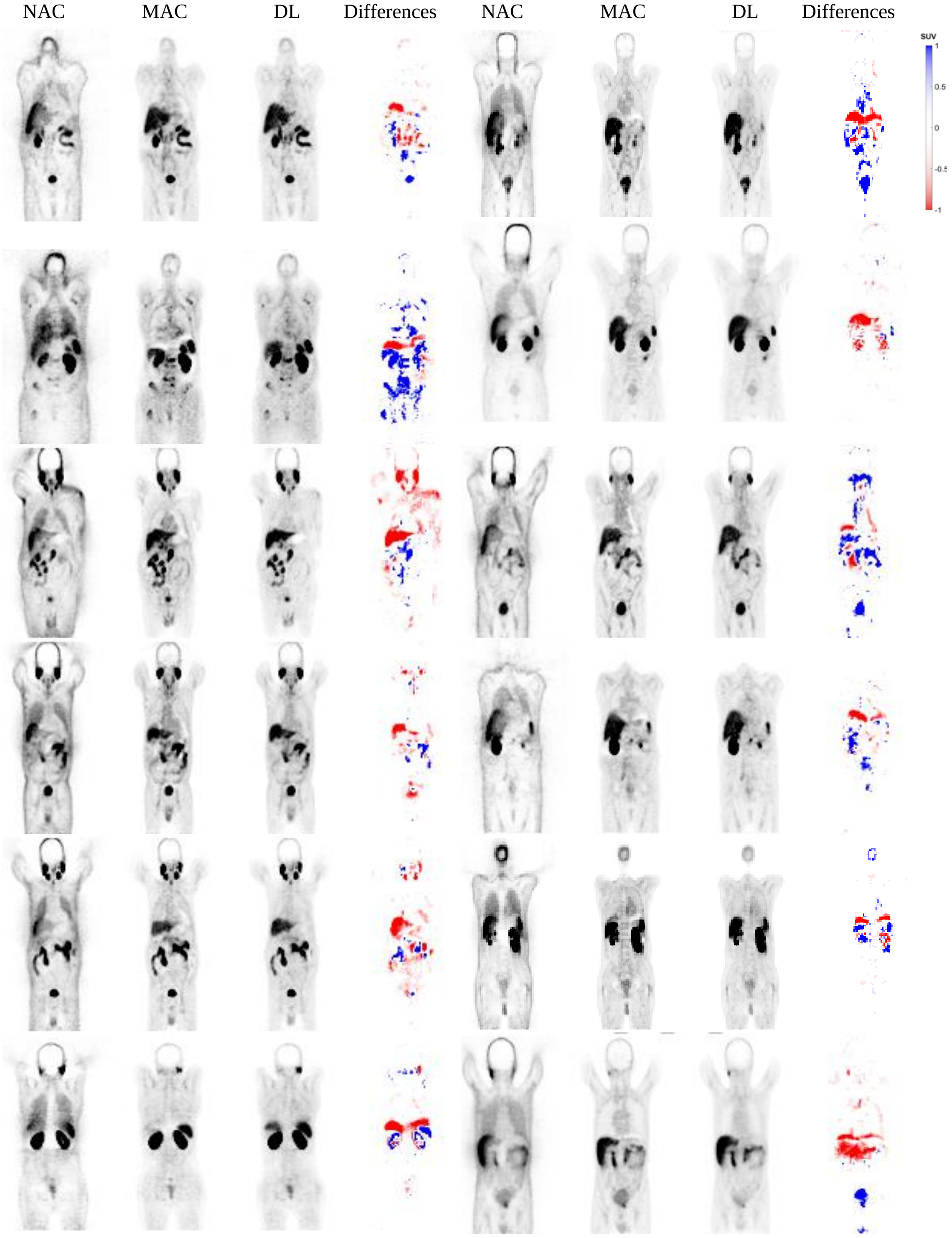


Figure 17: Coronal views of 12 clinical studies showing from left to right NAC, MAC, IMCM-DL, and the difference images of MAC and DL image. The images generated using the IMCM approach successfully corrected the mismatch artifact in the diaphragm, lung, liver, and spleen regions. The red intensity region represents areas where the DL image contains more information compared to the MAC. These regions are artifactual parts of the MAC images.

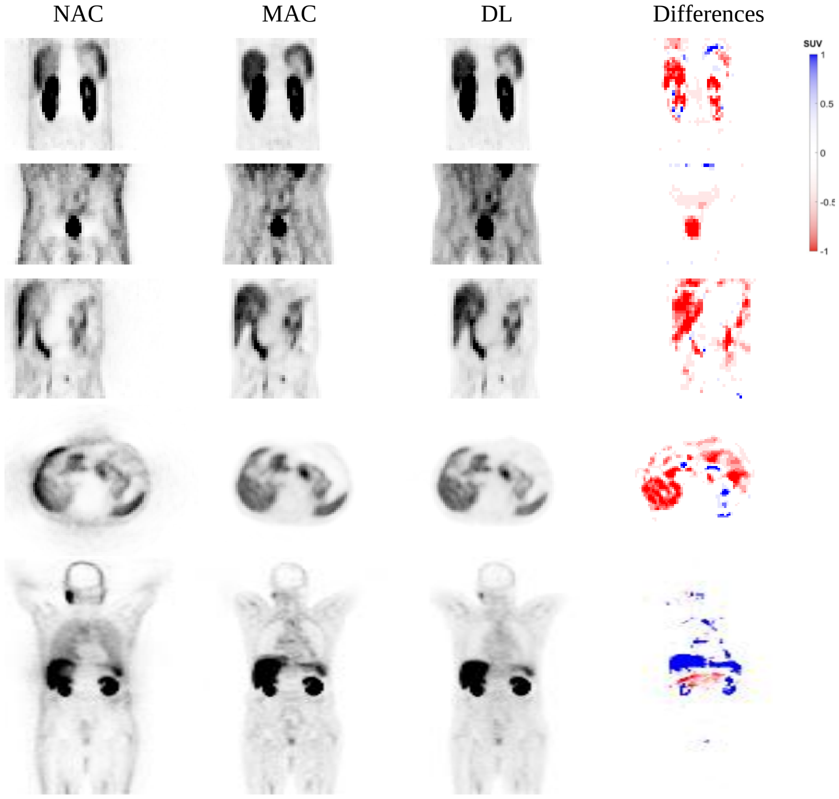


Figure 18: Coronal and axial views showing from left to right NAC, MAC, IMCM-DL and the difference images of MAC and DL image. The repeated scan which was requested right after the initial scan. The IMCM image recovered high quality and high diagnostic confidence for both scans. The red intensity region represents areas where the DL image contains more information compared to the MAC. These regions are artifactual parts of the MAC images.

Discussion

Various DL-based attenuation scatter correction (DL-ASC) methods have been developed for PET imaging (31,52,54,77,84,85,89,90). These include indirect approaches that generate attenuation maps from MRI or CT images (91,92). For instance, studies have employed GANs to achieve pseudo-CT images from NAC PET scans in both brain and whole-body PET imaging (50,51,76,89,93,94). Furthermore, the MLAA algorithm has been improved by incorporating DL to mitigate common issues such as crosstalk artifacts, slow convergence, and noisy attenuation maps (45,95,96). Direct DL-ASC methods bypass traditional methods by making ASC PET images directly from NAC images. This was first used in brain PET imaging and then tested in ^{18}F-FDG PET studies (51).

A significant challenge arises with the low tracer activity and the extensive positron range of ^{68}Ga-labelled pharmaceuticals, which generally produce lower-quality images compared to ^{18}F-labelled compounds (27,97). Initially, employing DL for direct ASC in PET might seem overly reliant on advanced technology (28,98–100). However, our findings indicate that it enhances both quantitative and qualitative aspects of PET images and effectively identifies and corrects mismatches and halo artifacts without needing anatomical images. While indirect techniques require reconstructions to produce ASC PET images, they often fail to address halo artifacts that arise during the reconstruction phase and are predominantly influenced by the PET images themselves.

In this study, we developed and evaluated the CT-free DL models for ASC in PET imaging. The IMCM and ADCM were tested across different centers and radiotracer distributions. The IMCM demonstrated better performance in terms of lower error metrics and higher similarity index compared to ADCM in dealing with external centers. Additionally, the tuned version of IMCM (TL-MC) showed acceptable adaptability and accuracy across different radiotracers.

The ADCM output has demonstrated that a single universal model may not be effective due to variations in tracer-injected activity across different hospitals. There is a need to tune radiotracer-wise models using heterogeneous datasets to address these discrepancies. However, using large and heterogeneous datasets from different hospitals in the same tracer can compensate for the differences in equipment, image acquisition, and reconstruction strategies. In our research, we utilized different data from various hospitals, which enhanced the accuracy of ASC in PET images when implementing a shared model across different hospitals for identical radiotracer imaging.

Furthermore, we employed the IMCM for additional qualitative analysis. Through quantitative assessments, we observed the substantial impact that radiotracers and scanners have on model performance. Notably, IMCM greatly increased the quantitative accuracy across various scanners, indicating the need for model tuning using transfer learning that is tailored to specific tracer situations and thus performs better than ADCM. IMCM showed enhanced efficiency when different scanners utilized the same radiotracer, compared to when various radiotracers were employed on the same scanner. We also found that the source of the data, including the type of scanner and radiotracer used, significantly affected ADCM's effectiveness, contrary to initial assumptions.

While the ADCM method focuses on decomposing the PET image correction process into separating anatomical and radiotracer-dependent information, our investigation couldn’t prove that this method is able to handle the differences between variant scanners and radiotracers well. This underscores the need for more robust, adaptable models like IMCM, which not only accommodate but thrive on the heterogeneity inherent in multi-center clinical data.

Normalization was therefore needed to ensure that the data was quantitatively comparable while being computationally straightforward for the DL approach. In this way by scaling the intensity values, one could easily rescale the images back to the original, which is vital for an accurate diagnosis and assessment of metabolic activity in a PET image. This had to be selected as a practical factor. We picked these factors for the ^{68}Ga and ^{18}FDG datasets based on publications (27,28), but regarding choosing the correct factor for ADCM, there was no reference. Since we might have extremely low voxel intensity in the denominator of Eq2, we saw high values for ADCM that may bias the model, such as outliers with values of 28180 and 7300, which were removed to align the focus on the representative range critical for analysis. Afterward, voxel intensities were normalized using a factor of 50 for relative, comparable, and manageable training values.

The joint histogram analysis raised questions regarding the calibration and reliability of the ADCM method in clinical situations. Notably, the overestimations observed by ADCM, especially in cross-center measurements (Figure 11), could lead to incorrect diagnoses in conditions where the accuracy of the SUV estimation is critical. The systematic bias towards higher SUV values, while giving a superficial appearance of accuracy as a higher R^2, suggests underlying problems in the algorithm or its application across different PET systems.

In contrast, the IMCM method's outcome with lower regression slopes, higher correlation coefficients, and more reliable SUV estimations, particularly in internal centers, showed its application in the clinic. The variance between IMCM and ADCM's performance shows the necessity for rigorous validation of imaging algorithms to ensure uniform performance across different settings. The analysis across cross-tracer highlights the critical aspect that a higher slope does not necessarily equate to better correlation. Instead, the consistency with which predictions align with actual values, as measured by correlation coefficients, provides a more substantial indication of a model's effectiveness. Despite lower regression slopes, the TL-MC model demonstrates a more reliable and consistent performance in capturing the true behavior of SUVs across the studied centers. Even the visualization comparison between IMCM and TL-MC (Figure 19) shows how important it is to tune the model specifically for each tracer's specific properties. This will make the model more useful and accurate in various clinical settings.

CT-ASCs are a primary adjustment for quantitative ^{68}Ga PET imaging. However, this process can introduce mismatches and halo artifacts in ^{68}Ga PET images, potentially altering patient diagnosis and prognosis (27,28,36). These artifacts are challenging to detect and correct in real clinical settings. Our developed model does not require image reconstruction with ASC. Our results align well with the findings reported in the previous study (27). Notably, our approach demonstrates a significant advantage by effectively handling external test sets without the need for transfer learning in the case of the same radiotracer. The qualitative analysis demonstrated the effectiveness of our proposed model in detecting and removing mismatches and halo artifacts in the chest, abdomen, and pelvic regions without needing ground truth in ^{68}Ga PET images. We also observed scenarios where repeated scans, typically conducted to eliminate artifacts, failed and even exacerbated them (Figure 20). Here, our DL algorithms could distinguish and correct these issues independently of the ground truth.

Previous studies' predominant limitation lies in their single-center datasets, which restrict the generalizability of DL models (27,84,85). Our current study employs a multi-center approach to address this issue. Future research should explore clinical imaging parameters such as SUV_{mean}, SUV_{max}, and total lesion metabolism, providing a more comprehensive analysis of the IMCM model's performance. These metrics, along with an assessment of the most relevant radiomic features within the sphere of influence, will provide crucial insights into the model's effectiveness under various clinical conditions.

Future investigations should focus on the performance of the IMCM model in organ-specific evaluations for both clean and artifactual images. Such targeted analysis would provide a more specific understanding of the model's capabilities in different clinical situations. In addition, rigorous statistical testing of categorized outcomes will be important. In the future, other models could be investigated since we developed this image-to-image translation model based on a segmentation model. Different types of neural networks might be used, such as generative adversarial networks, variational autoencoders, transformer-based models, or Swin-UNet.

Moreover, the changing of the model hyperparameters would further improve the performance. Any other optimizers, Adaptive Moment Estimation with Weight Decay (AdamW) (101,102), Root Mean Square Propagation (RMSprop) (103), or Nesterov-accelerated Adaptive Moment Estimation (Nadam) (104), might be more efficient than the classical ones. With tools like Optuna (105), checking hyperparameters can systemically find an optimal setting and possibly even enhance accuracy and the improvement of its robustness in general.

The model performance should be evaluated on data from different centers to guarantee that it will generalize and be robust across varying clinical environments. These tests will provide deeper insights into the model's consistency and reliability across different diagnostic categories and will help refine the model's application and improve diagnostic accuracy in practical healthcare settings.

Conclusion

In this thesis, we have demonstrated the efficacy of an Integrated multi-center Dynamic Unet DL framework for artifact detection and correction in PET imaging of ^{68}Ga-labelled compounds. The approach leverages large datasets from multiple centers. Through the incorporation of transfer learning concepts, we have developed site-specific models that significantly outperform those based on single-center data, thereby addressing a major limitation in the field of medical imaging. Our model effectively detected and corrected artifacts. This enhancement is vital for making therapeutic decisions in the field of oncology, where PET imaging plays a central role in diagnosing, planning treatments, and evaluating responses. By using Dyn-Unet architecture and other advanced DL techniques, our method has not only improved image quality but also greatly decreased the appearance of common artifacts like halo and mismatch artifacts, especially in ^{68}Ga-PET imaging. The effective implementation of our models in different centers highlights their resilience and flexibility, which are essential for general acceptance in clinical settings.

Declaration

I acknowledge the use of OpenAI's ChatGPT for assistance in rewriting initial sentences to be more professional and grammatically correct. All content has been reviewed and edited to ensure accuracy and relevance.

Code availability

The source code is available on GitHub with this link. Please note that this repository is private, and access is restricted to authorized individuals only.

References

1. Cerqueira MD. Cardiac SPECT or PET?: Is there still a debate? Vol. 29, Journal of Nuclear Cardiology. 2022.

2. Sarikaya I. Cardiac applications of PET. Nucl Med Commun [Internet]. 2015 Oct;36(10):971–85. Available from: https://journals.lww.com/00006231-201510000-00002

3. Catana C, Procissi D, Wu Y, Judenhofer MS, Qi J, Pichler BJ, et al. Simultaneous in vivo positron emission tomography and magnetic resonance imaging. Proc Natl Acad Sci U S A. 2008;105(10).

4. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Vol. 42, European Journal of Nuclear Medicine and Molecular Imaging. 2015.

5. Karakatsanis NA, Fokou E, Tsoumpas C. Dosage optimization in positron emission tomography: state-of-the-art methods and future prospects. Am J Nucl Med Mol Imaging. 2015;5(5).

6. Fahey FH, Treves ST, Adelstein SJ. Minimizing and communicating radiation risk in pediatric nuclear medicine. J Nucl Med Technol. 2012;40(1).

7. Zaidi H, MML. Scatter Compensation Techniques in PET. PET clinics. PET Clin [Internet]. 2007 [cited 2023 Nov 20];2(2):219–34. Available from: https://doi.org/10.1016/j.cpet.2007.10.003

8. Baer M, Kachelrie M. Hybrid scatter correction for CT imaging. Phys Med Biol. 2012;57(21).

9. Watson CC, Casey ME, Michel C, Bendriem B. Advances in scatter correction for 3D PET/CT. In: IEEE Nuclear Science Symposium Conference Record. 2004.

10. Zaidi H, Koral KF. Scatter modelling and compensation in emission tomography. Vol. 31, European Journal of Nuclear Medicine and Molecular Imaging. 2004.

11. Pettinato C, Nanni C, Farsad M, Castellucci P, Sarnelli A, Civollani S, et al. Artefacts of PET/CT images. Biomed Imaging Interv J. 2006;2(4).

12. Lammertsma AA. Forward to the past: The case for quantitative PET imaging. Vol. 58, Journal of Nuclear Medicine. 2017.

13. Hasegawa BH, Gingold EL, Reilly SM, Liew SC, Cann CE. Description of a simultaneous emission-transmission CT system. In: Medical Imaging IV: Image Formation. 1990.

14. Presotto L, Busnardo E, Perani D, Gianolli L, Gilardi MC, Bettinardi V. Simultaneous reconstruction of attenuation and activity in cardiac PET can remove CT misalignment artifacts. Journal of Nuclear Cardiology. 2016;23(5).

15. Mostafapour S, Greuter M, van Snick JH, Brouwers AH, Dierckx RAJO, van Sluis J, et al. Ultra-low dose CT scanning for PET/CT. Med Phys. 2024;51(1).

16. Beyer T, Townsend DW, Brun T, Kinahan PE, Charron M, Roddy R, et al. A combined PET/CT scanner for clinical oncology. Journal of Nuclear Medicine. 2000;41(8).

17. Townsend DW. Physical principles and technology of clinical PET imaging. Vol. 33, Annals of the Academy of Medicine Singapore. 2004.

18. Kinahan PE, Townsend DW, Beyer T, Sashin D. Attenuation correction for a combined 3D PET/CT scanner. Med Phys. 1998;25(10).

19. Kuttner S, Lassen ML, Øen SK, Sundset R, Beyer T, Eikenes L. Quantitative PET/MR imaging of lung cancer in the presence of artifacts in the MR-based attenuation correction maps. Acta radiol. 2020;61(1).

20. Wagenknecht G, Kaiser HJ, Mottaghy FM, Herzog H. MRI for attenuation correction in PET: Methods and challenges. Vol. 26, Magnetic Resonance Materials in Physics, Biology and Medicine. 2013.

21. Hofmann M, Pichler B, Schölkopf B, Beyer T. Towards quantitative PET/MRI: A review of MR-based attenuation correction techniques. Eur J Nucl Med Mol Imaging. 2009;36(SUPPL. 1).

22. Catana C, Van Der Kouwe A, Benner T, Michel CJ, Hamm M, Fenchel M, et al. Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype. Journal of Nuclear Medicine. 2010;51(9).

23. Keereman V, Mollet P, Berker Y, Schulz V, Vandenberghe S. Challenges and current methods for attenuation correction in PET/MR. Vol. 26, Magnetic Resonance Materials in Physics, Biology and Medicine. 2013.

24. Martinez-Moller A, Souvatzoglou M, Delso G, Bundschuh RA, Chefd’Hotel C, Ziegler SI, et al. Tissue classification as a potential approach for attenuation correction in whole-body PET/MRI: Evaluation with PET/CT data. Journal of Nuclear Medicine. 2009;50(4).

25. Sureshbabu W, Mawlawi O. PET/CT Imaging Artifacts* [Internet]. Vol. 33, J Nucl Med Technol. 2005. Available from: http://www.snm.org/ce_online

26. Mawlawi O, Pan T, Macapinlac HA. PET/CT Imaging Techniques, Considerations, and Artifacts. J Thorac Imaging [Internet]. 2006;21(2). Available from: https://journals.lww.com/thoracicimaging/fulltext/2006/05000/pet_ct_imaging_techniques,_considerations,_and.2.aspx

27. Shiri I, Salimi Y, Maghsudi M, Jenabi E, Harsini S, Razeghi B, et al. Differential privacy preserved federated transfer learning for multi-institutional 68Ga-PET image artefact detection and disentanglement. Eur J Nucl Med Mol Imaging. 2023;

28. Shiri I, Salimi Y, Hervier E, Pezzoni A, Sanaat A, Mostafaei S, et al. Artificial Intelligence-Driven Single-Shot PET Image Artifact Detection and Disentanglement: Toward Routine Clinical Image Quality Assurance. Clin Nucl Med. 2023 Dec 1;48(12):1035–46.

29. Abdoli M, Dierckx RAJO, Zaidi H. Metal artifact reduction strategies for improved attenuation correction in hybrid PET/CT imaging. Vol. 39, Medical Physics. 2012.

30. Ghafarian P, Aghamiri SMR, Ay MR, Rahmim A, Schindler TH, Ratib O, et al. Is metal artefact reduction mandatory in cardiac PET/CT imaging in the presence of pacemaker and implantable cardioverter defibrillator leads? Eur J Nucl Med Mol Imaging. 2011;38(2).

31. Lindemann ME, Nensa F, Quick HH. Impact of improved attenuation correction on 18F-FDG PET/MR hybrid imaging of the heart. PLoS One. 2019;14(3).

32. McQuaid SJ, Hutton BF. Sources of attenuation-correction artefacts in cardiac PET/CT and SPECT/CT. Eur J Nucl Med Mol Imaging. 2008;35(6).

33. Fendler WP, Schmidt DF, Wenter V, Thierfelder KM, Zach C, Stief C, et al. 68Ga-PSMA PET/CT detects the location and extent of primary prostate cancer. Journal of Nuclear Medicine. 2016;57(11).

34. Afshar-Oromieh A, Hetzheim H, Kratochwil C, Benesova M, Eder M, Neels OC, et al. The theranostic PSMA ligand PSMA-617 in the diagnosis of prostate cancer by PET/CT: Biodistribution in humans, radiation dosimetry, and first evaluation of tumor lesions. Journal of Nuclear Medicine. 2015;56(11).

35. Rauscher I, Maurer T, Beer AJ, Graner FP, Haller B, Weirich G, et al. Value of 68Ga-PSMA HBED-CC PET for the assessment of lymph node metastases in prostate cancer patients with biochemical recurrence: Comparison with histopathology after salvage lymphadenectomy. Journal of Nuclear Medicine. 2016;57(11).

36. Heußer T, Mann P, Rank CM, Schäfer M, Dimitrakopoulou-Strauss A, Schlemmer HP, et al. Investigation of the halo-artifact in 68Ga-PSMA-11-PET/MRI. PLoS One. 2017;12(8).

37. Magota K, Numata N, Shinyama D, Katahata J, Munakata Y, Maniawski PJ, et al. Halo artifacts of indwelling urinary catheter by inaccurate scatter correction in 18F-FDG PET/CT imaging: incidence, mechanism, and solutions. EJNMMI Phys. 2020;7(1).

38. Fourquet A, Lahmi L, Rusu T, Belkacemi Y, Créhange G, de la Taille A, et al. Restaging the biochemical recurrence of prostate cancer with [68Ga]Ga-PSMA-11 PET/CT: Diagnostic performance and impact on patient disease management. Cancers (Basel). 2021;13(7).

39. Dinges J, Nekolla SG, Bundschuh RA. Motion artifacts in oncological and cardiac PET imaging. Vol. 8, PET Clinics. 2013.

40. Presotto L. The long fight against motion artifacts in cardiac PET. Vol. 29, Journal of Nuclear Cardiology. 2022.

41. Piccinelli M, Votaw JR, Garcia E V. Motion Correction and Its Impact on Absolute Myocardial Blood Flow Measures with PET. Vol. 20, Current Cardiology Reports. 2018.

42. Mehranian A, Zaidi H. Joint Estimation of Activity and Attenuation in Whole-Body TOF PET/MRI Using Constrained Gaussian Mixture Models. IEEE Trans Med Imaging. 2015;34(9).

43. Chun SY, Kim KY, Lee JS, Fessier JA. Joint estimation of activity distribution and attenuation map for TOF-PET using alternating direction method of multiplier. In: Proceedings - International Symposium on Biomedical Imaging. 2016.

44. Mehranian A, Arabi H, Zaidi H. Vision 20/20: Magnetic resonance imaging-guided attenuation correction in PET/MRI: Challenges, solutions, and opportunities. Med Phys. 2016;43(3).

45. Li S, Wang G. Modified kernel MLAA using autoencoder for PET-enabled dual-energy CT. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2021;379(2204).

46. Carney JPJ, Townsend DW, Rappoport V, Bendriem B. Method for transforming CT images for attenuation correction in PET/CT imaging. Med Phys. 2006;33(4).

47. Alessio AM, Kohlmyer S, Branch K, Chen G, Caldwell J, Kinahan P. Cine CT for attenuation correction in cardiac PET/CT. Journal of Nuclear Medicine. 2007;48(5).

48. Alberts I, Hünermund JN, Prenosil G, Mingels C, Bohn KP, Viscione M, et al. Clinical performance of long axial field of view PET/CT: a head-to-head intra-individual comparison of the Biograph Vision Quadra with the Biograph Vision PET/CT. Eur J Nucl Med Mol Imaging. 2021;48(8).

49. Guo R, Xue S, Hu J, Sari H, Mingels C, Zeimpekis K, et al. Using domain knowledge for robust and generalizable deep learning-based CT-free PET attenuation and scatter correction. Nat Commun. 2022 Dec 1;13(1).

50. Yang J, Sohn JH, Behr SC, Gullberg GT, Seo Y. Ct-less direct correction of attenuation and scatter in the image space using deep learning for whole-body fdg pet: Potential benefits and pitfalls. Radiol Artif Intell. 2021 Mar 1;3(2).

51. Shiri I, Ghafarian P, Geramifar P, Leung KHY, Ghelichoghli M, Oveisi M, et al. Direct attenuation correction of brain PET images using only emission data via a deep convolutional encoder-decoder (Deep-DAC). Eur Radiol. 2019 Dec 1;29(12):6867–79.

52. Lee JS. A Review of Deep-Learning-Based Approaches for Attenuation Correction in Positron Emission Tomography. Vol. 5, IEEE Transactions on Radiation and Plasma Medical Sciences. 2021.

53. Qian H, Rui X, Ahn S. Deep Learning Models for PET Scatter Estimations. In: 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). 2017. p. 1–5.

54. Liu F, Jang H, Kijowski R, Zhao G, Bradshaw T, McMillan AB. A deep learning approach for18 f-fdg pet attenuation correction. EJNMMI Phys. 2018;5(1).

55. Wu X, Sahoo D, Hoi SCH. Recent advances in deep learning for object detection. Neurocomputing. 2020;396.

56. Zhao ZQ, Zheng P, Xu ST, Wu X. Object Detection with Deep Learning: A Review. Vol. 30, IEEE Transactions on Neural Networks and Learning Systems. 2019.

57. Ma X, Wu J, Xue S, Yang J, Zhou C, Sheng QZ, et al. A Comprehensive Survey on Graph Anomaly Detection With Deep Learning. IEEE Trans Knowl Data Eng. 2023;35(12).

58. McLeavy CM, Chunara MH, Gravell RJ, Rauf A, Cushnie A, Staley Talbot C, et al. The future of CT: deep learning reconstruction. Vol. 76, Clinical Radiology. 2021.

59. Ahishakiye E, Van Gijzen MB, Tumwiine J, Wario R, Obungoloch J. A survey on deep learning in medical image reconstruction. Vol. 1, Intelligent Medicine. 2021.

60. Kim SH, Choi YH, Lee JS, Lee SB, Cho YJ, Lee SH, et al. Deep learning reconstruction in pediatric brain MRI: comparison of image quality with conventional T2-weighted MRI. Neuroradiology. 2023;65(1).

61. Jebur RS, Zabil MHBM, Hammood DA, Cheng LK. A comprehensive review of image denoising in deep learning. Multimed Tools Appl. 2023;

62. Tian C, Fei L, Zheng W, Xu Y, Zuo W, Lin CW. Deep learning on image denoising: An overview. Vol. 131, Neural Networks. 2020.

63. Wu H, Liu Y, Wang J. Review of text classification methods on deep learning. Vol. 63, Computers, Materials and Continua. 2020.

64. Ibrahim DM, Elshennawy NM, Sarhan AM. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. Comput Biol Med. 2021;132.

65. Krishna MM, Neelima M, Harshali M, Rao MVG. Image classification using Deep learning. International Journal of Engineering and Technology(UAE). 2018;7.

66. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability (Switzerland). 2021;13(3).

67. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. IET Image Process. 2022;16(5).

68. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image Segmentation Using Deep Learning: A Survey. IEEE Trans Pattern Anal Mach Intell. 2022;44(7).

69. Xia T, Alessio AM, Kinahan PE. Limits of ultra-low dose CT attenuation correction for PET/CT. In: IEEE Nuclear Science Symposium Conference Record. 2009.

70. Prieto E, García-Velloso MJ, Aquerreta JD, Rosales JJ, Bastidas JF, Soriano I, et al. Ultra-low dose whole-body CT for attenuation correction in a dual tracer PET/CT protocol for multiple myeloma. Physica Medica. 2021;84.

71. Wafa B, Moussaoui A. A review on methods to estimate a CT from MRI data in the context of MRI-alone RT. Medical Technologies Journal. 2018;2(1).

72. Lindemann ME, Gratz M, Blumhagen JO, Jakoby B, Quick HH. MR-based truncation correction using an advanced HUGE method to improve attenuation correction in PET/MR imaging of obese patients. Med Phys. 2022;49(2).

73. Sun H, Xi Q, Fan R, Sun J, Xie K, Ni X, et al. Synthesis of pseudo-CT images from pelvic MRI images based on an MD-CycleGAN model for radiotherapy. Phys Med Biol. 2022;67(3).

74. Wang T, Manohar N, Lei Y, Dhabaan A, Shu HK, Liu T, et al. MRI-based treatment planning for brain stereotactic radiosurgery: Dosimetric validation of a learning-based pseudo-CT generation method. Medical Dosimetry. 2019;44(3).

75. Jabbarpour A, Mahdavi SR, Vafaei Sadr A, Esmaili G, Shiri I, Zaidi H. Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and CycleGAN: Dosimetric assessment for 3D conformal radiotherapy. Comput Biol Med. 2022;143.

76. Shiri I, Arabi H, Geramifar P, Hajianfar G, Ghafarian P, Rahmim A, et al. Deep-JASC: joint attenuation and scatter correction in whole-body 18F-FDG PET using a deep residual network. Eur J Nucl Med Mol Imaging. 2020 Oct 1;47(11):2533–48.

77. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging-based attenuation correction for PET/MR imaging. Radiology. 2018;286(2).

78. Arabi H, Zaidi H. Deep learning–based metal artefact reduction in PET/CT imaging. Eur Radiol. 2021;31(8).

79. Arabi H, Zaidi H. Truncation compensation and metallic dental implant artefact reduction in PET/MRI attenuation correction using deep learning-based object completion. Phys Med Biol. 2020;65(19).

80. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Vol. 25, Nature Medicine. 2019.

81. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises. Proceedings of the IEEE. 2021;109(5).

82. Scott IA, Cook D, Coiera EW, Richards B. Machine learning in clinical practice: prospects and pitfalls. Medical Journal of Australia. 2019;211(5).

83. Saboury B, Bradshaw T, Boellaard R, Buvat I, Dutta J, Hatt M, et al. Artificial Intelligence in Nuclear Medicine: Opportunities, Challenges, and Responsibilities Toward a Trustworthy Ecosystem. Journal of Nuclear Medicine. 2023;64(2).

84. Shiri I, Vafaei Sadr A, Akhavan A, Salimi Y, Sanaat A, Amini M, et al. Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning. Eur J Nucl Med Mol Imaging. 2023 Mar 1;50(4):1034–50.

85. Shiri I, Sadr A V, Sanaat A, Ferdowsi S, Arabi H, Zaidi H. Federated Learning-based Deep Learning Model for PET Attenuation and Scatter Correction: A Multi-Center Study. In: 2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). 2021. p. 1–3.

86. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. In: Informatik aktuell. 2019.

87. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: Breaking the Spell on Successful Medical Image Segmentation Fabian. Nat Methods. 2021;18(2).

88. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. 2004;13(4).

89. Hwang D, Kang SK, Kim KY, Seo S, Paeng JC, Lee DS, et al. Generation of PET attenuation map for whole-body time-of-flight 18F-FDG PET/MRI using a deep neural network trained with simultaneously reconstructed activity and attenuation maps. Journal of Nuclear Medicine. 2019;60(8).

90. McMillan AB, Bradshaw TJ. Artificial Intelligence–Based Data Corrections for Attenuation and Scatter in Position Emission Tomography and Single-Photon Emission Computed Tomography. Vol. 16, PET Clinics. W.B. Saunders; 2021. p. 543–52.

91. Arabi H, Zaidi H. Magnetic resonance imaging-guided attenuation correction in whole-body PET/MRI using a sorted atlas approach. Med Image Anal. 2016;31.

92. Akbarzadeh A, Ay MR, Ahmadian A, Riahi Alam N, Zaidi H. MRI-guided attenuation correction in whole-body PET/MR: Assessment of the effect of bone attenuation. Ann Nucl Med. 2013;27(2).

93. Armanious K, Hepp T, Küstner T, Dittmann H, Nikolaou K, La Fougère C, et al. Independent attenuation correction of whole body [18F]FDG-PET using a deep learning approach with Generative Adversarial Networks. EJNMMI Res. 2020;10(1).

94. Izadi S, Shiri I, F. Uribe C, Geramifar P, Zaidi H, Rahmim A, et al. Enhanced direct joint attenuation and scatter correction of whole-body PET images via context-aware deep networks. Z Med Phys. 2024;

95. Shi L, Zhang J, Toyonaga T, Shao D, Onofrey JA, Lu Y. Deep learning-based attenuation map generation with simultaneously reconstructed PET activity and attenuation and low-dose application. Phys Med Biol. 2023;68(3).

96. Hwang D, Kim KY, Kang SK, Seo S, Paeng JC, Lee DS, et al. Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning. Journal of Nuclear Medicine. 2018;59(10).

97. Hong I, Nekolla SG, Michel C. Improving Scatter Correction for Ga-68 PSMA PET Studies. In: 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2017 - Conference Proceedings. 2018.

98. Shiri I, Salimi Y, Sanaat A, Saberi A, Amini M, Akhavanallaf A, et al. Fully Automated PET Image Artifacts Detection and Correction Using Deep Neural Networks &lt;/strong&gt; Journal of Nuclear Medicine [Internet]. 2022 Jun 1;63(supplement 2):3218. Available from: http://jnm.snmjournals.org/content/63/supplement_2/3218.abstract

99. Shiri I, Sanaat A, Salimi Y, Akhavanallaf A, Arabi H, Rahmim A, et al. PET-QA-Net: Towards Routine PET Image Artifact Detection and Correction using Deep Convolutional Neural Networks. In: 2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). 2021. p. 1–3.

100. Afshar-Oromieh A, Wolf M, Haberkorn U, Kachelrieß M, Gnirs R, Kopka K, et al. Effects of arm truncation on the appearance of the halo artifact in 68Ga-PSMA-11 (HBED-CC) PET/MRI. Eur J Nucl Med Mol Imaging. 2017;44(10).

101. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019. 2019.

102. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.

103. Tieleman, T., & Hinton G. Divide the gradient by a running average of its recent magnitude. Human and Machine Hearing. 2012;4(2).

104. Dozat T. Incorporating Nesterov Momentum into Adam. ICLR Workshop. 2016;(1).

105. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.

106. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016.

107. He K, Zhang X, Ren S, Sun J. U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015;

Supplementary Material 1

The initial Step from the Segmentation task to the image-to-image translation

The basic idea is that if a DL model can accurately identify and position organs for segmentation, then it could supposedly learn to correct an image in the desired style by using the correct activation functions, loss functions, and an appropriate architectural design. So, for the first step, we tried to find the answer to this question: “Could a model trained on arbitrary images learn to produce an acceptable output by using the same image as both input and target?” We first focused on visual acceptability rather than quantitative metrics. We utilized CT images as samples before accessing the original data. The experimental setup involved using these images as both the training inputs and target inputs, aiming to fine-tune the model’s hyperparameters to achieve visually satisfactory outputs. This stage was primarily about understanding the influence of various parameters on the initial results and was not concerned with the precision of error metrics. Fig 1 in this supplementary section illustrates some of the outputs. This stage served an educational purpose, helping us to understand the foundational dynamics of DL applications in corrected images.

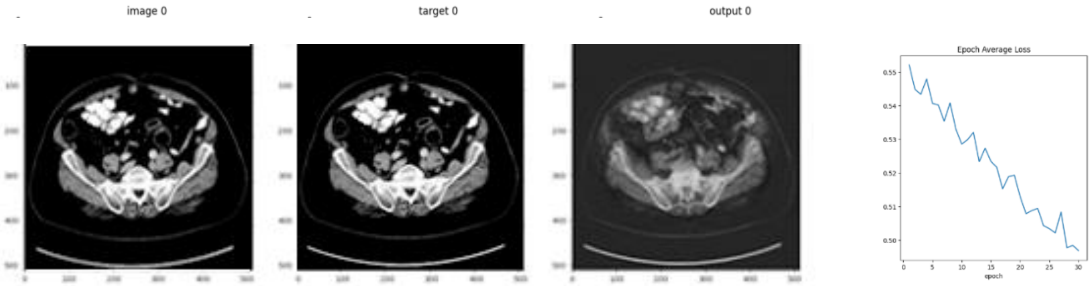


Fig 1: One slice of output and raining loss, from the left to right: input, target and output of the model.

Different Models

3D-Unet-Model

Following the initial phase, we progressed to applying the developed model to the ^{68}Ga dataset. To adapt the model for our dataset, several transformations and optimization of hyperparameters tuned to better process the specific profiles of ^{68}Ga images. First, we checked the model for just one patient's data. Fig 2 this section shows the outputs from this phase of the project.

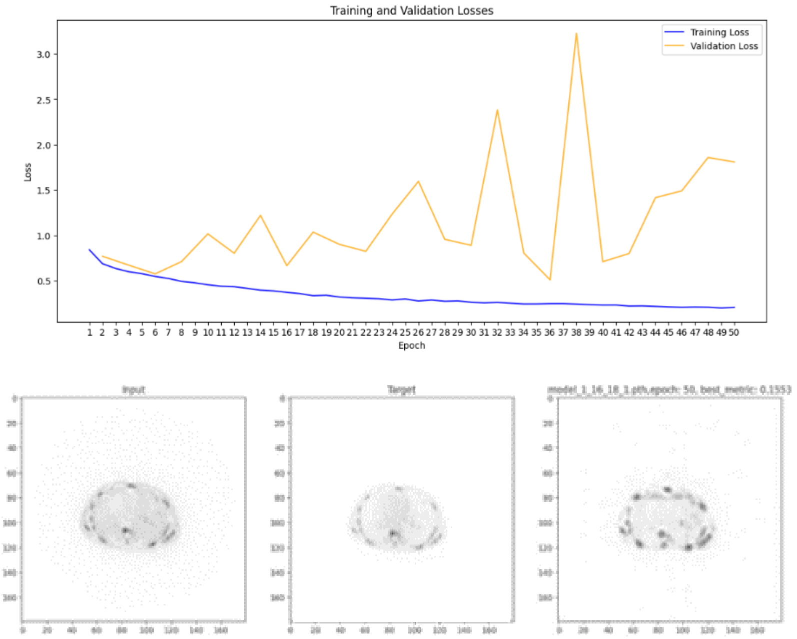


Fig 2: top: Training and validation loss for the 3D-Unet model, bottom: One slice of output. And then we tried it for a portion of the data (20 patients)

It is obvious that there was still some patch pattern on the image, which means there are parameters that need to be changed. Fig 3 concluded after adapting the spacing, dimensions, and other parameters for loading the data appropriate for our dataset and using all datasets.

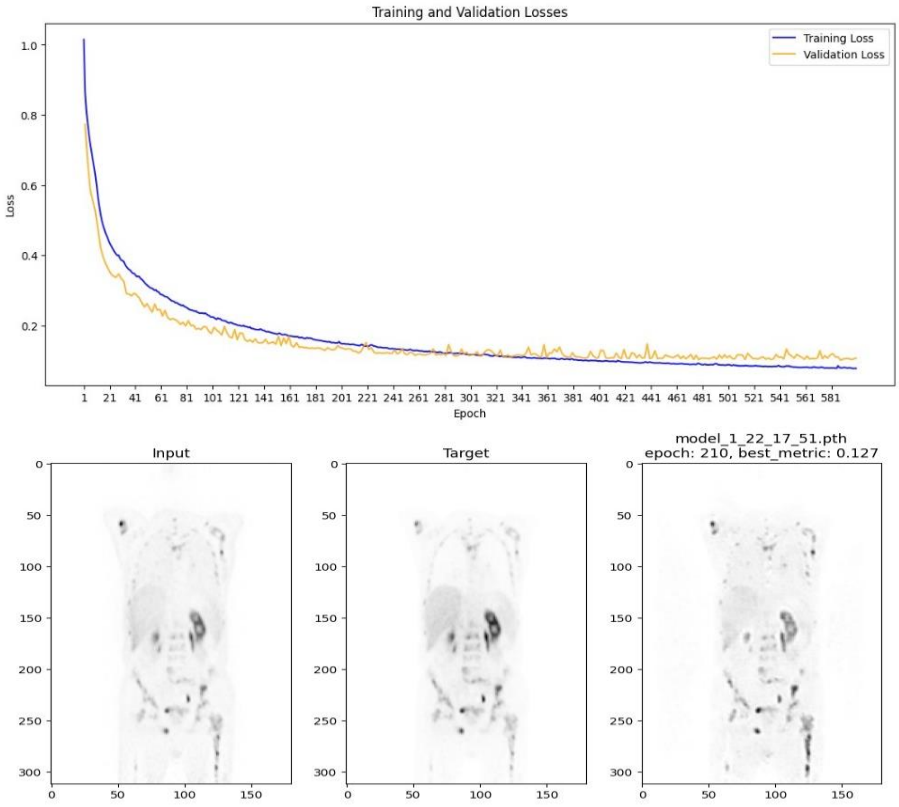


Fig 3: top: Training and validation loss for 3D-Unet model, bottom: One slice of output.

Patched-3D U-net:

In the initial phase of our research, we attempted to use full-body 3D PET data as single inputs for training our DL model (106). This approach, however, presented significant challenges. The limited number of available data and the limited computational resources required to process full-body 3D data. Most researchers in this field typically use a 2D slice-wise approach using data-frame images, which significantly reduces the computational demand. Others utilize a smaller section of the data frame, training their models patch-wise to manage resource constraints effectively. Considering these factors, we opted to focus on using image patches exclusively in the axial direction and fixed boundaries in the coronal and sagittal dimensions 168 and 168, with each patch containing 32 axial slices. This approach effectively increased our data tenfold, facilitating more extensive training under limited resource conditions. The outcomes of this method are presented in Fig 4. These results underline the adaptability of our approach in optimizing data usage and computational resources while still enabling robust model training.

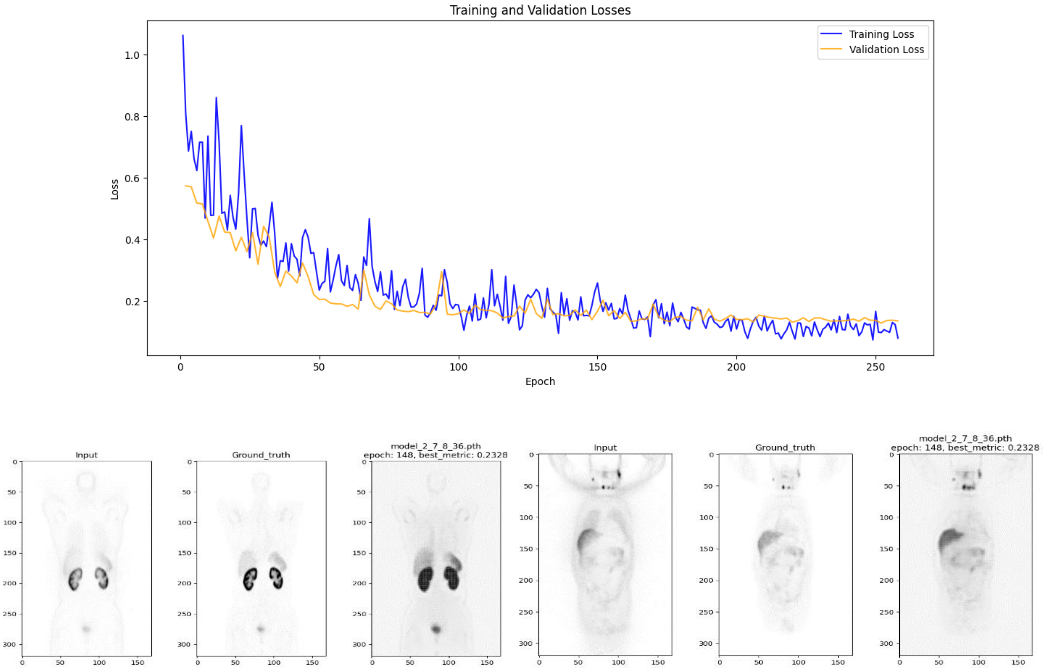


Fig 4: top: Training and validation loss for 3D-Unet model, bottom: Two sample slices of outputs, Best Metric: 0.2328, Epoch: 148

2D-Unet

In addition to searching for the best model to get lower loss and better image quality, we evaluated a 2D-Unet model training approach (107). This 2D U-Net architecture was mostly like the previous model. The model training was optimized using an Adam optimizer with a specifically customized plan learning rate, which adjusted the learning rate based on the epoch count to enhance training stability and performance. Some key variables and results are detailed in Fig 5.

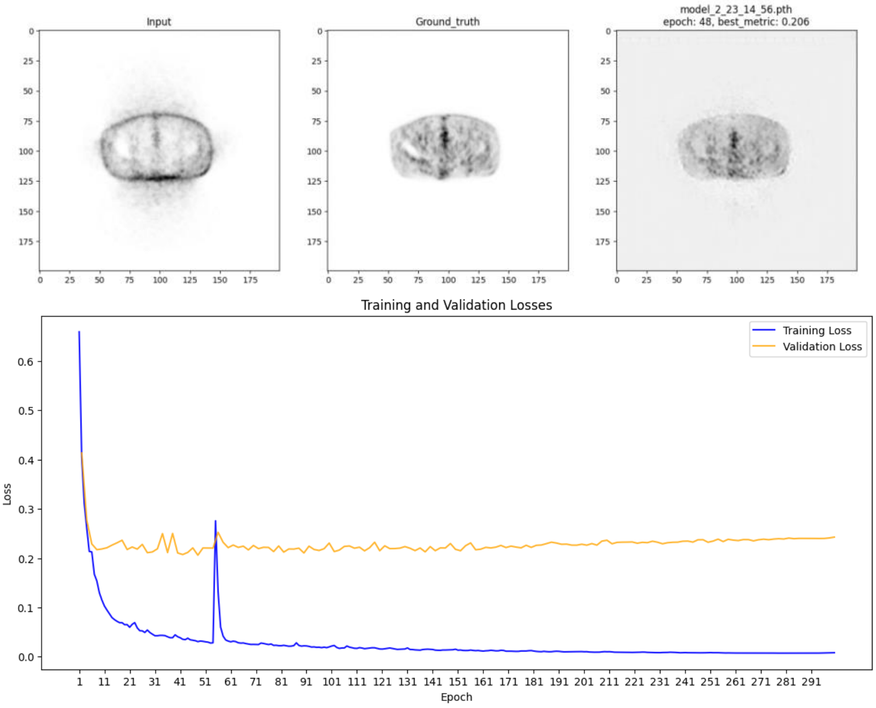


Fig 5: top: Training and validation loss for 2D-Unet model, bottom: Sample slice of output, Best Metric: 0.206, Epoch: 48

DyUnet:

In parallel with 2D evaluation, we implemented the DynUNet architecture (86,87), an advanced and dynamic variant of the traditional U-Net designed specifically for biomedical image segmentation. DynUNet introduces several key enhancements over the standard U-Net, including the option for deep supervision. This feature allows the network to output additional intermediate layers' predictions and facilitate the learning process by ensuring that gradients are effectively propagated back through the network, enhancing the training dynamics and enabling the model to learn detailed representations without significant overfitting. With compatible configurations of kernel sizes, strides, and depth of architecture, models enable effective capture of relevant features at different scales. Key configuration parameters of DynUNet are listed in Table 2, and there is one sample output from our initial implementation in Fig 6.

Table 3: Some specifications of training approach

patch_size	[168, 168, 16]
spacing	[4.07, 4.07, 3.00]
spatial_size	(168, 168, 320)
train_transforms	Spacingd(keys=["image", "target"], pixdim= spacing, mode= 'trilinear'),SpatialPadd(keys=["image","target"], spatial_size=spatial_size, mode='constant'),RandSpatialCropSamplesd(keys=["image","target"], roi_size=self.patch_size, num_samples=4),
val_transforms	CenterSpatialCropd(keys=["image", "target"], roi_size=self.spatial_size)
Model	DynUNet(spatial_dims=3, in_channels=1, out_channels=1, kernel_size=kernels, strides=strides, upsample_kernel_size=strides[1:], norm_name="INSTANCE", deep_supervision=True, deep_supr_num=2,)

After these improvements, we could finally decrease the validation loss from around 0.2 at the initial trials to 0.0664. To enhance the robustness of our model, we implemented specific data augmentations. These included adding rotations of ±15 degrees and increasing the number of samples per patient from 4 to 20.

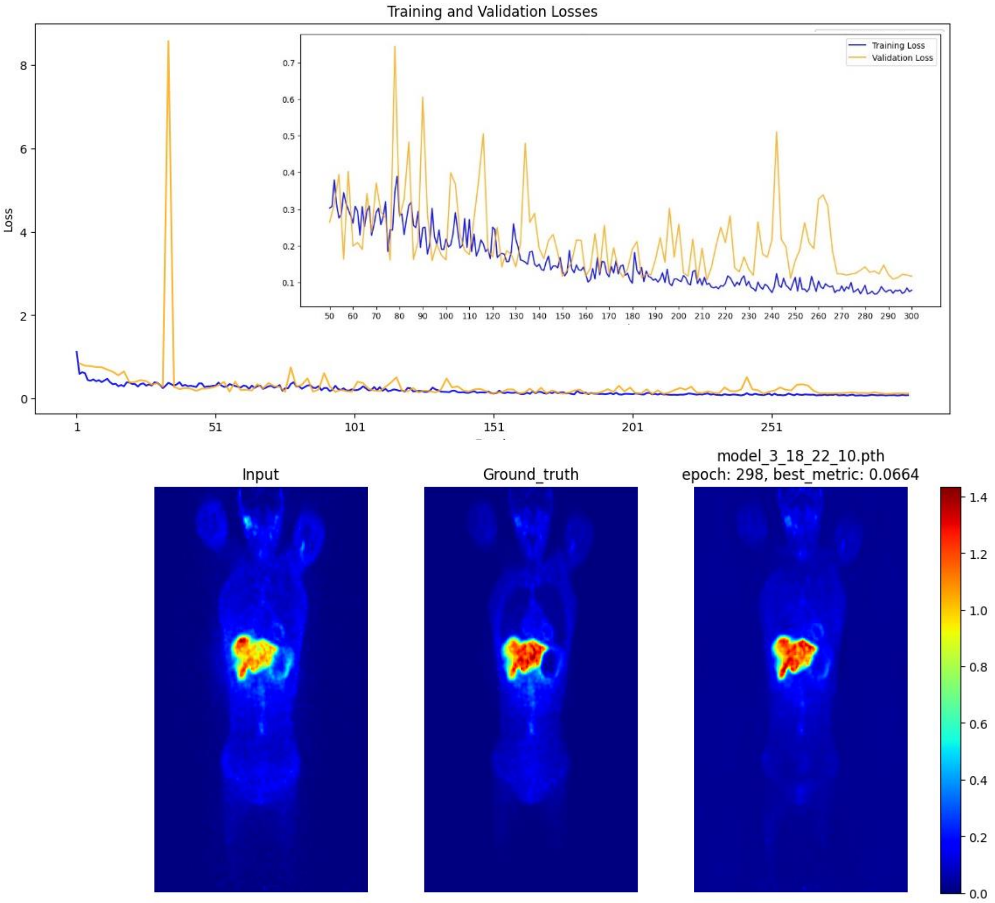


Fig 6: top: Training and validation loss for 2D-Unet model, bottom: Sample slice of output, Best Metric: 0.206, Epoch: 48

Here are some other metric errors from the beginning of this research:

At early stage	ME: 0.64 MAE: 0.95 RE: 193.7% ARE: 199.0%
Using Unet	Mean Error (SUV): -0.32 ± 0.1032 Mean Absolute Error (SUV): 0.33 ± 0.0868 Relative Error (SUV%): -55.49 ± 15.6193 Absolure Relative Error (SUV%): 56.98 ± 13.3306 Root Mean Squared Error: 0.48 ± 0.1741 Peak Signal-to-Noise Ratio: 23.92 ± 6.4356 Structual Similarity Index: 0.63 ± 0.1537
Using DynUnet	mean_error: -0.43 ± 0.3433 mean_absolute_error: 0.54 ± 0.2896 relative_error: -23.92 ± 14.8091 absolute_relative_error: 35.36 ± 7.7831 rmse: 1.13 ± 0.8008 psnr: 32.57 ± 4.2616 ssim: 0.87 ± 0.0568
DynUnet, ADCM method	Mean Error (SUV): -0.42 ± 0.0783 Mean Absolute Error (SUV): 0.42 ± 0.0767 Relative Error (SUV%): -72.41 ± 10.2247 Absolure Relative Error (SUV%): 72.65 ± 9.9125 Root Mean Squared Error: 0.57 ± 0.1856 Peak Signal-to-Noise Ratio: 22.53 ± 6.7792 Structual Similarity Index: 0.44 ± 0.1617

Supplementary Material 2

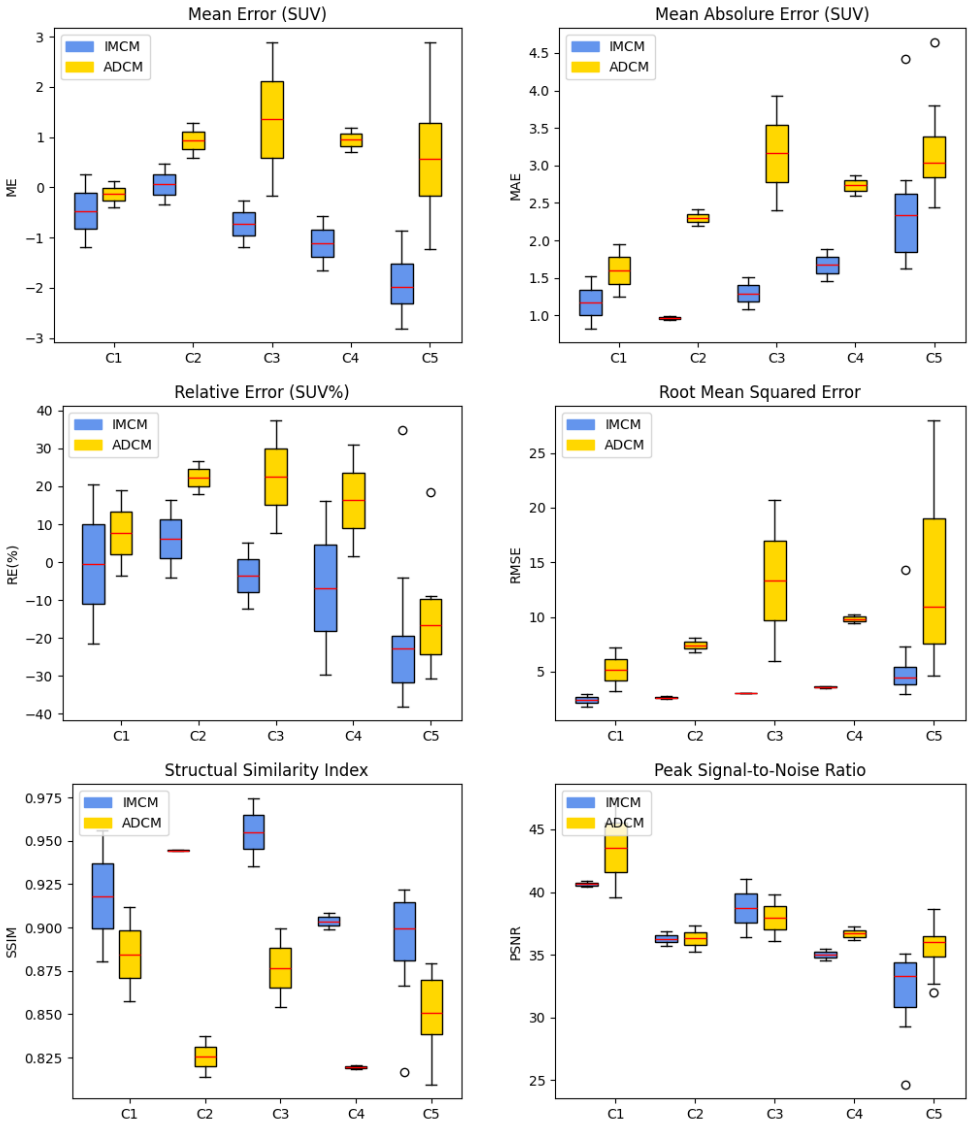


Fig 7: Performance Metrics of IMCM and ADCM Across Centers C1 to C5

Table 4: Summary statistics of quantitative parameters for different approaches on cross center (Ga dataset)

	Method	ME	MAE	RE	ARE	RMSE	PSNR	SSIM
Mean ± SD	**ADCM**	0.67 ± 1.10	2.87 ± 0.75	-2.17 ± 20.85	57.23 ± 7.41	11.79 ± 7.03	36.83 ± 3.17	0.85 ± 0.03
	IMCM	-1.38 ± 0.93	1.94 ± 0.83	-12.38 ± 20.98	43.62 ± 11.56	4.40 ± 2.66	34.42 ± 3.92	0.91 ± 0.04
CI95%	**ADCM**	[0.15, 1.18]	[2.52, 3.22]	[-11.93, 7.59]	[53.77, 60.70]	[8.50, 15.08]	[35.35, 38.31]	[0.84, 0.86]
	IMCM	[-1.81, -0.94]	[1.55, 2.33]	[-22.20, -2.56]	[38.21, 49.04]	[3.16, 5.65]	[32.58, 36.25]	[0.89, 0.92]

Statistical tests

Normality Testing

Before selecting an appropriate statistical test for our analysis, we first assessed the normality of the distribution of each metric within both datasets using the Shapiro-Wilk test. This step was crucial to determine whether parametric or non-parametric statistical methods were suitable. Our findings indicated that several metrics did not follow a normal distribution, particularly in the IMCM dataset, where metrics such as Relative Error (SUV%) and Absolute Relative Error (SUV%) showed significant deviations from normality with p-values below 0.05. Similarly, Root Mean Squared Error and Peak Signal-to-Noise Ratio in the ADCM dataset also deviated significantly from a normal distribution.

Table 5: Evaluation of normality of all metric variables across both ADCM and IMCM datasets by performing a Shapiro-Wilk test for each metric.

	Metric	ADCM Statistic	ADCM P-value	IMCM Statistic	IMCM P-value
0	Mean Error (SUV)	0.962684	0.598745	0.964505	0.637189
1	Mean Absolute Error (SUV)	0.973161	0.819726	0.902938	0.046832
2	Relative Error (SUV%)	0.926644	0.133062	0.903215	0.047397
3	Absolute Relative Error (SUV%)	0.934748	0.190480	0.813324	0.001375
4	Root Mean Squared Error	0.875041	0.014425	0.670732	0.000018
5	Peak Signal-to-Noise Ratio	0.826691	0.002222	0.944862	0.295736
6	Structural Similarity Index	0.963606	0.618108	0.973200	0.820480

Choice of Statistical Test

Given the non-normality observed in several key metrics across the datasets, we used the Wilcoxon signed-rank test, a non-parametric method, for our analysis. This test is particularly advantageous as it does not assume the normality of the data and is ideal for comparing two related samples or repeated measurements on a single sample. This was chosen to handle the paired nature of our data, where each center was analyzed under both ADCM and IMCM conditions.

Our analysis revealed significant differences in several metrics between the ADCM and IMCM methodologies. Notably, the Mean Error (SUV) and Absolute Relative Error (SUV%) showed considerable variations, suggesting distinct impacts of the two methodologies on these particular metrics. The Wilcoxon test results indicated statistically significant differences with low p-values, underscoring the effectiveness of one method over the other in specific conditions.

Table 6: Summarized results of the Wilcoxon test with the False Discovery Rate (FDR) corrections applied to the p-values.

Metric	U-statistic	P-value
Mean Error (SUV)	371.0	0.000039
Mean Absolute Error (SUV)	330.0	0.000460
Relative Error (SUV%)	267.0	0.072045
Absolute Relative Error (SUV%)	357.0	0.000023
Root Mean Squared Error	364.0	0.000097
Peak Signal-to-Noise Ratio	286.0	0.020734
Structural Similarity Index	42.0	0.000020

The results from the Wilcoxon test show statistically significant differences between the ADCM and IMCM datasets for most of the image-derived metrics, except for the "Relative Error (SUV%)," for which the corrected p-value does not indicate a statistically significant difference.

Table 7: Summary statistics of quantitative parameters for different approaches on cross tracer (FDG dataset)

	Method	**ME**	**MAE**	**RE**	**ARE**	**RMSE**	**PSNR**	**SSI**
Mean ± SD	**ADCM**	0.29 ± 0.58	1.08 ± 0.35	34.08 ± 48.96	80.22 ± 34.25	3.71 ± 4.14	37.38 ± 3.89	0.77 ± 0.09
	TL-MC	-0.54 ± 0.15	0.69 ± 0.13	-39.70 ± 9.13	52.11 ± 7.61	1.18 ± 0.61	35.27 ± 6.18	0.78 ± 0.11
CI95%	**ADCM**	[0.02, 0.55]	[0.92, 1.24]	[11.80, 56.37]	[64.63, 95.81]	[1.82, 5.59]	[35.61, 39.15]	[0.72, 0.81]
	TL-MC	[-0.60, -0.47]	[0.63, 0.75]	[-43.86, -35.55]	[48.64, 55.57]	[0.91, 1.46]	[32.46, 38.09]	[0.73, 0.83]

Table 8: Summary statistics of quantitative parameters for different centers tuned for each radiotracer separately (TL-MC) and tested on all test sets (centers 1-7).

Quantitative metric	**Center 1-4**	**Center 5**	**Center 6**	**Center 7**	**All Test Set**
ME	-0.56 ± 0.74	-1.92 ± 0.58	-0.46 ± 0.16	-0.61 ± 0.09	-0.95 ± 0.78
MAE	1.28 ± 0.37	2.38 ± 0.76	0.64 ± 0.13	0.73 ± 0.12	1.30 ± 0.86
RE	-1.15 ± 18.77	-19.87 ± 19.58	-35.66 ± 11.69	-43.38 ± 3.55	-26.38 ± 21.03
ARE	36.38 ± 7.12	48.45 ± 11.62	49.56 ± 8.11	54.42 ± 6.64	47.97 ± 10.53
RMSE	2.90 ± 0.58	5.41 ± 3.05	1.00 ± 0.25	1.35 ± 0.78	2.75 ± 2.49
PSNR	37.66 ± 2.67	32.25 ± 3.04	37.74 ± 6.59	33.03 ± 5.07	34.86 ± 5.16
SSIM	0.93 ± 0.03	0.89 ± 0.03	0.80 ± 0.13	0.76 ± 0.092	0.84 ± 0.11
CI 95%					
ME	[-1.18, 0.06]	[-2.29, -1.56]	[-0.57, -0.34]	[-0.67, -0.55]	[-1.19, -0.70]
MAE	[0.97, 1.59]	[1.90, 2.87]	[0.55, 0.73]	[0.65, 0.81]	[1.03, 1.57]
RE	[-16.84, 14.55]	[-32.31, -7.43]	[-44.02, -27.29]	[-45.77, -41.00]	[-33.01, -19.74]
ARE	[30.42, 42.34]	[41.07, 55.84]	[43.75, 55.37]	[49.96, 58.89]	[44.65, 51.29]
RMSE	[2.42, 3.38]	[3.47, 7.35]	[0.82, 1.18]	[0.83, 1.88]	[1.97, 3.54]
PSNR	[35.43, 39.90]	[30.32, 34.18]	[37.62, 37.85]	32.97 to 33.09	[33.23, 36.48]
SSIM	[0.90, 0.96]	[0.87, 0.91]	0.68 to 0.91	0.70 to 0.82	[0.81, 0.87]

Column “Center 1-4” represents the results of testing on the whole test set when training is performed on the center 1 to 4 data set. “Center 5” represents an external center with the same radiotracer, and Center 6 & 7 test sets represent the results of tuned models, in which training and testing are performed for different radiotracers (whole 20% of the clean dataset).