Medical Image Synthesis from CT to PET using Convolutional Neural Network

Xiaoyu Deng[1], Kouki Nagamune[1,2], and Hiroki Takada[1]

[1]University of Fukui, 3-9-1 Bunkyo, Fukui, 910-0019, Japan
[2]University of Hyogo, 2167 Shosha, Himeji, Hyogo, 670-2280, Japan

Abstract

U-Net, a deep learning architecture, has gained widespread application in medical image processing due to its exceptional performance and efficient structural design. This architecture enhances traditional convolutional neural networks with its symmetric "U" shape, making it extensively utilized for applications such as image denoising, medical image registration, attenuation correction, lesion segmentation, and facial image restoration. This study leverages the U-Net architecture for cross-modality conversion from computed tomography to positron emission tomography images. Preliminary results demonstrate rapid performance improvements within the initial training epochs, achieving stability and high-quality reconstruction as training progresses. Despite observable fluctuations in performance metrics, which highlight the model's challenges in generalizing across the inherent variability of medical imaging datasets, the U-Net model exhibits robustness in image reconstruction. With further adjustments and optimizations, there is potential for enhanced performance in future applications, promising advances in the practical application of deep learning techniques in medical imaging.

1 Introduce

U-Net is a deep learning architecture initially designed for medical image segmentation tasks and has become widely popular in the field of medical image processing due to its outstanding performance and efficient structure. The architecture of U-Net improves upon the conventional convolutional neural network (CNN) by featuring a symmetric "U" shape, including a contracting path (encoder) and an expanding path (decoder). A key characteristic of U-Net is the use of skip connections, which connect feature maps from the contracting path to corresponding layers in the expanding path. This design enables the network to utilize more precise spatial information during the upsampling process, which is crucial for enhancing the accuracy of segmentation boundaries, particularly when precise delineation of structures in medical images is essential.

Common strategies for increasing the scale of the U-Net model include deepening the network, expanding the number of feature map channels, refining the skip connection structure, and incorporating attention modules such as Transformer-based architectures. While these improvements can enhance model performance, they also increase model complexity and computational requirements, often encountering performance bottlenecks. In this study, we employ multiple simple encoder-decoder structures to construct a generative network and evaluate the performance of multi-stage models in medical image generation tasks.

2 Related Works

Since Olaf and colleagues [1] introduced U-Net, it has become widely used due to its structural advantages and excellent performance in applications such as image denoising, medical image registration, and attenuation correction. U-Net has also been applied to various other image segmentation tasks, including lesion segmentation and facial image restoration.

Armanious et al. [2] proposed an end-to-end medical image-to-image translation framework using GAN, demonstrating its effectiveness in tasks like PET-CT translation, MR motion artifact correction, and PET image denoising.

Singh et al. [3] presented an automated medical image registration method based on U-Net, using GAN to generate pseudo-CT images from non-attenuation corrected PET images, improving the efficiency and accuracy of coronary angiography registration.

Liu et al. [4] developed a method for generating pseudo-CT images used for attenuation correction from single non-attenuation corrected 18F-FDG

PET images.

Du et al. [5] reviewed six U-Net-based methods applied in medical image segmentation, including segmentation tasks for pulmonary nodules, the heart, and the brain.

Zeng et al. [6] demonstrated good performance using a two-stage cascaded U-Net in facial image restoration tasks, suggesting that multi-stage cascaded U-Nets hold potential advantages in image generation tasks.

Singh and Liu et al. applied U-Net architectures by fine-tuning modules for medical image registration and attenuation correction, achieving promising results in specialized domains.

While Armanious and Zeng et al. employed cascaded U-Net structures, they did not explore the performance of multi-stage cascaded U-Nets in medical image generation tasks.

This research utilizes a multi-stage cascaded U-Net model for CT-to-PET image translation and evaluates model performance based on the following metrics.

3 Method

U-Net's architecture is designed symmetrically, with a contracting path (Encoder) to capture context and a symmetric expanding path (Decoder) that enables precise localization. This study aims to construct a standard U-Net, a convolutional neural network, to develop a cross-modality medical image converter from CT to PET images. The model will be optimized using a specific loss function suitable for this type of image conversion task.

3.1 Encoder

The encoder follows the typical architecture of a convolutional network. It consists of repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. The convolution operation in U-Net can be described by the following equation:

$$I' = \sum_{i,j} (I * K)(i,j) + b$$

where I represents the input image, K is the convolution kernel, b is the bias, and I' is the output feature map.

Max pooling in the encoder is used to reduce the spatial dimensions of the input feature maps:

$$P(x,y) = \max(I(x+i, y+j))$$

where $P(x,y)$ is the pooled output, and $I(x+i, y+j)$ represents the input in the neighborhood around location (x,y).

3.2 Decoder

The decoder includes a series of upsampling and convolution operations. Each step in the expanding path includes an upsampling of the feature map followed by a 2×2 convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU. Up sampling in the expanding path uses transposed convolutions to increase the size of the feature map:

$$U = K^T * I$$

where K^T is the transposed convolution kernel and U is the upsampled output.

4 Experiments

This study employs the U-Net architecture for cross-modality medical image conversion tasks, specifically to construct a U-Net that inputs a CT image and converts it into a corresponding PET image. In this research, the lung PET or CT scan data were powered by the National Cancer Institute Cancer Imaging Program (CIP) [7]. The dataset encompasses 251,135 lung scan images from 355 subjects, primarily collected between 2009 and 2011, including each subject's gender, age, weight, smoking history, and cancer diagnosis classification. All scan data in the dataset are stored in DICOM format. This study processed these 251,135 scan data using the MicroDicom software on a Windows operating system. The subjects in the dataset are labeled according to the type of cancer: Type A for adenocarcinoma, Type B for small cell carcinoma, Type E for large cell carcinoma, and Type G for squamous cell carcinoma. Not all subjects' data include both PET and CT scans. Therefore, this study selected only the scan data of 38 subjects diagnosed with small cell carcinoma (Type B), which include PET scans, various CT scans, and fusion-enhanced scan images. Among these 38 subjects, only 9 had both PET and CT scans, totaling 12,930 scan images. Through precise selection, PET/CT scans with a slice position error of no more

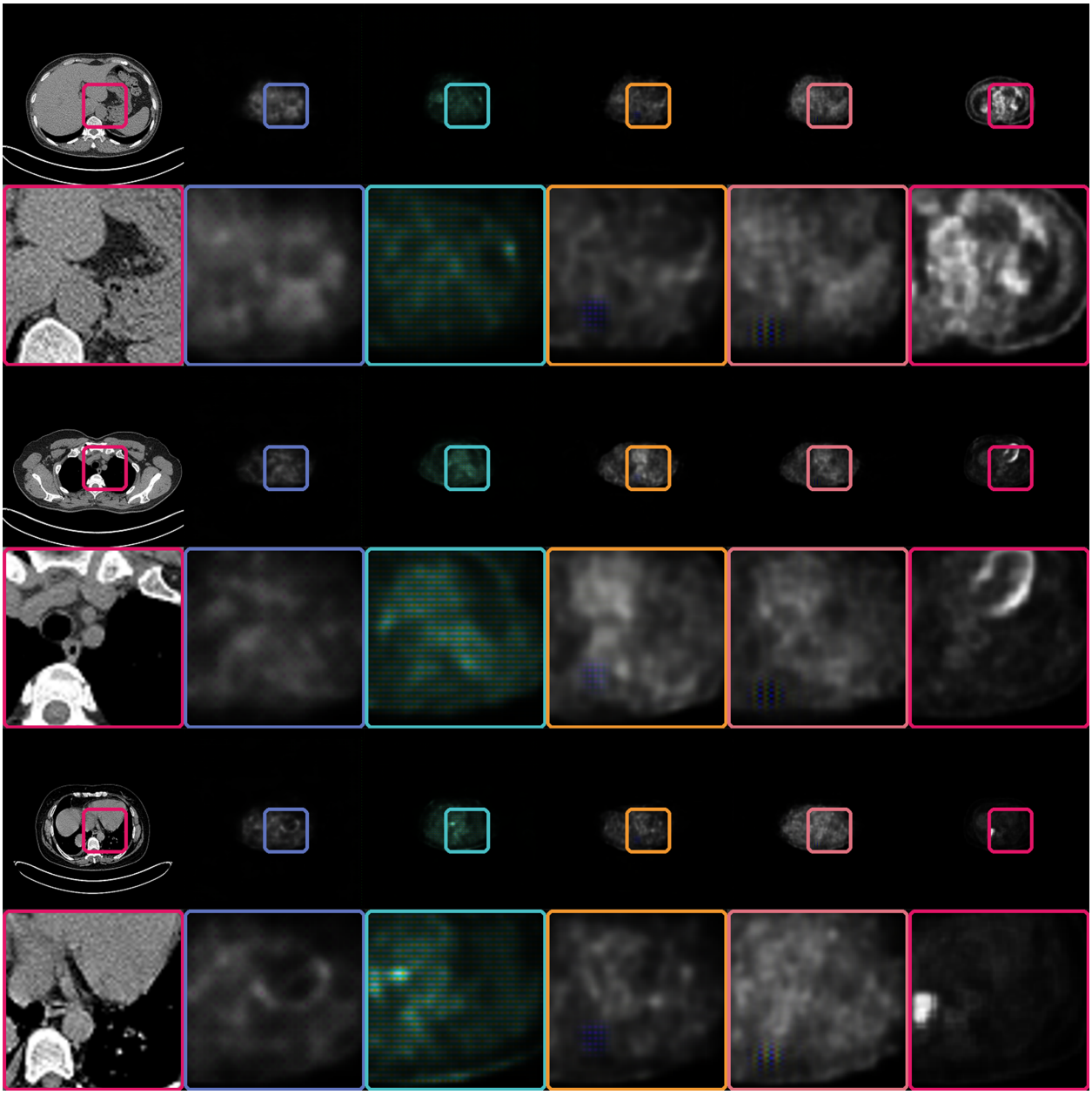

Figure 1: Gennerated PET Image and Real CT PET Image

than 0.2 mm were defined as paired scan data, ultimately obtaining 928 scan images. These 464 pairs of PET/CT lung scan images are used for this study.

In this study, a standard U-Net model was constructed, with a parameter scale of approximately 50 million. We exported the obtained 464 pairs of PET/CT lung scan data as 256x256 pixel RGB format PNG images, which were independently divided into training and testing datasets. The specific division of the dataset is shown in the following table1.

Table 1: Dataset Partition of Experiment

Params count	Test	Train	Total
Lung PET/CT Pair	64	400	464
Total Images	128	800	928

In this experiment, the standard U-Net model is employed to synthesize positron emission tomography (PET) images from computed tomography (CT) images as inputs. Only the mean squared error (MSE) loss is used as the loss function, capitalizing on the

benefits of L2 loss. The model optimizer used is the Adam optimizer, with a learning rate (lr) set at 0.001. This relatively low learning rate assists the model in steadily approaching the global optimum during the training process. The best experimental results are shown in Table **??**.

Table 2: SSIM,PSNR,MSE Results of Experiment

Stage Count	SSIM	PSNR	MAE	MSE
1 Stage	-	-	-	-
2 Stages	-	-	-	-
3 Stages	-	-	-	-
4 Stages	-	-	-	-
5 Stages	-	-	-	-
6 Stages	-	-	-	-

To enhance the model's generalization ability and mitigate overfitting issues, L2 regularization is employed with a weight decay parameter set at 0.001. The experiment records data over 200 training epochs, including Loss, SSIM, multi-scale SSIM, VIF, PSNR, and MSE. Tests are conducted after each epoch, with changes in the Loss, training and testing SSIM, MS-SSIM, VIF, PSNR, and MSE metrics as indicated in the respective tables. The model demonstrates commendable performance on both the training and testing datasets.

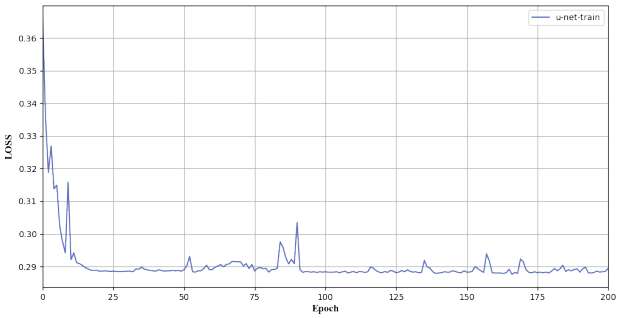

Figure 2: Loss Line Figure of All Epoch in Train Process

Figure 2 shows that during the initial few training epochs, the loss rapidly decreases from approximately 0.36 to close to 0.29. This indicates that the model quickly improved its fit to the training data, making the learning process highly effective during this stage. After the rapid decline, the loss values tend to stabilize, fluctuating around 0.29. This trend suggests that the model may have reached a stable state in its training, where further learning yields limited performance improvements. Although the loss values are relatively stable, there are still minor fluctuations at certain epochs, such as around the 50th, 100th, and 175th epochs. These fluctuations could be due to random factors in the training process, such as random selection of batch data, adjustments in the learning rate, or external disturbances. Overall, the loss values remain relatively low and stable over 200 training epochs, indicating that the model possesses a certain degree of generalization ability and has maintained good stability throughout the training process.

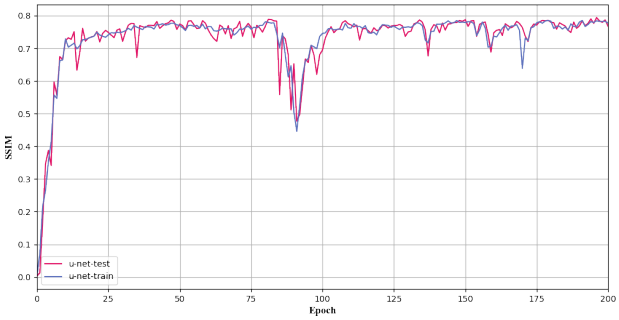

Figure 3: SSIM Line Figure of All Epoch in Train and Test Process

Figure 3 shows that the SSIM values rapidly increase during the initial training phase (first 25 epochs), rising from near zero to above 0.6. This indicates that the U-Net model quickly learned the mapping relationship between CT and PET images, showing significant initial learning effects. After this rapid growth, the SSIM values on both the training and testing sets tend to stabilize, mainly fluctuating around 0.7. This demonstrates that the model maintained high image reconstruction quality during the continued training process. Throughout the training period, the SSIM values on the training set are usually higher than those on the testing set, which might be a sign of slight overfitting, suggesting that the model may be overly optimized for the training data and slightly less generalized for unseen test data.

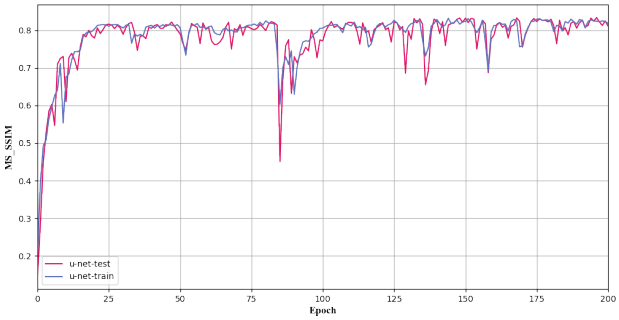

Figure 4: MS-SSIM Line Figure of All Epoch in Train and Test Process

Figure 4 illustrates that, for both the training and testing datasets, the MS-SSIM values rapidly improve within the initial few epochs. This shows that the model effectively captured the key features and structural information necessary for the conversion from CT to PET images early in the learning process. After the initial rapid improvement, the MS-SSIM values enter a phase of more stable fluctuations, as reflected in both the training and testing data. The MS-SSIM values for the training data are slightly higher than those for the testing data, a common phenomenon in machine learning models, which may indicate slight overfitting. After about 50 epochs, despite some fluctuations, the MS-SSIM values generally remain at a high level, especially above 0.7, indicating that the model has a good ability to capture the structural similarity of the images. The MS-SSIM values on the testing dataset are generally slightly lower than those on the training dataset, but the gap between them is small, indicating that the model also performs well on unseen data.

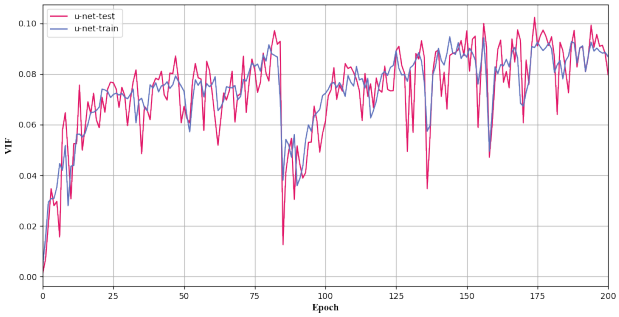

Figure 5: VIF Line Figure of All Epoch in Train and Test Process

As shown in Figure 5, during the initial few epochs, the Visual Information Fidelity (VIF) index rapidly increases, indicating that the model quickly adapts to the data, effectively reconstructing visual information from CT images into PET images. This is a positive sign, demonstrating the model's efficiency in learning and capturing key features for the conversion between the two image types. After the initial growth, the VIF values exhibit similar fluctuation patterns in both training and testing processes, likely due to variability encountered with different data batches. The figure reveals significant peaks and troughs at certain epochs, such as around the 75th, 100th, and 175th, which may relate to specific images or noisy images in the dataset. The VIF curves for training and testing generally converge throughout the experiment, although the testing curve is usually slightly lower than the training curve. This trend indicates that the model has some degree of generalization capability, although this capability may have limitations.

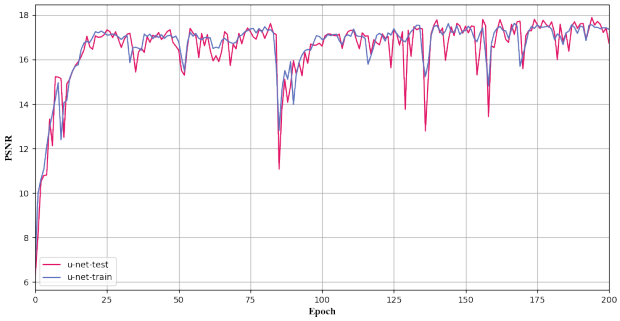

Figure 6: PSNR Line Figure of All Epoch in Train and Test Process

Figure 6 displays that, for both training (blue line) and testing (red line) sets, the Peak Signal-to-Noise Ratio (PSNR) values rapidly increase during the initial epochs, suggesting that the model quickly begins to effectively reconstruct PET images from CT images. The rapid improvement initially may be related to swift adjustments in the model parameters, significantly enhancing the quality of the reconstructed images. Following the initial growth, the PSNR values tend to stabilize but exhibit several significant fluctuations during both training and testing processes. These fluctuations might reflect changes in performance when the model encounters difficult samples within the data or due to optimization algorithms (such as learning rate adjustments). Despite these fluctuations, PSNR generally remains at a relatively high level, indicating that the model reconstructs PET images well and shows consistent performance throughout the training and testing periods. The PSNR curves for the training and testing sets are very close, indicating good generalization capability of the model to unseen data. Nonetheless, the testing curve occasionally dips below the training curve, which might be due to particularly challenging image samples in the test set.

From Figure 7, it can be observed that the Mean Squared Error (MSE) for both the training and testing sets rapidly decreases from initial values of around 0.20 to below 0.05 during the initial few epochs. This indicates that the model quickly and effectively adapts to the image conversion from CT to PET, with a rapid learning rate. After the quick decline, the training and testing MSE tend to stabilize, although there are several pronounced peaks, particularly around the 125th and 175th epochs. These peaks may reflect performance fluctuations when the model processes certain specific data batches or anomalous image samples. Besides occasional peaks,

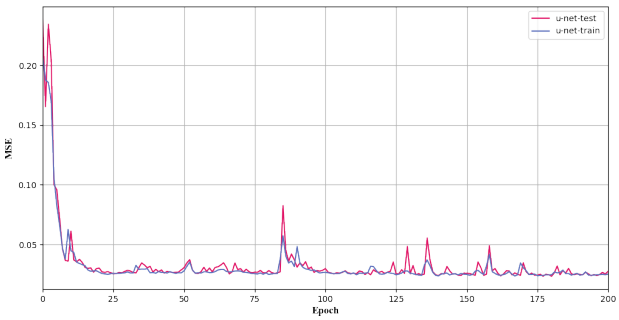

Figure 7: MSE Line Figure of All Epoch in Train and Test Process

MSE remains at a low level most of the time, showing relative stability of the model throughout the training and testing processes. This suggests that the model can reliably reconstruct images in most cases. The MSE curves for the training and testing sets are very close, indicating good generalization capability of the model to unseen data. This close trend also suggests that there are no significant issues with overfitting.

5 Conclusion

This research thoroughly explores the development history, technical principles, and practical applications of U-Net, and conducts an empirical study in medical image classification tasks. The experimental results demonstrate that the U-Net model shows high efficiency and stability in the task of converting CT to PET images. With further adjustments and optimizations, it is hoped that even better performance can be achieved in future applications.

Acknowledge

We would like to express our sincere gratitude to the National Cancer Institute Cancer Imaging Program for generously making their high-quality medical imaging dataset available and authorized for use on the Internet, providing indispensable resources for the smooth conduct of this research.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[2] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684, January 2020.

[3] Ananya Singh, Jacek Kwiecinski, Sebastien Cadet, Aditya Killekar, Evangelos Tzolos, Michelle C Williams, Marc R. Dweck, David E. Newby, Damini Dey, and Piotr J. Slomka. Automated nonlinear registration of coronary PET to CT angiography using pseudo-CT generated from PET with generative adversarial networks. *Journal of Nuclear Cardiology*, 30(2):604–615, April 2023.

[4] Fang Liu, Hyungseok Jang, Richard Kijowski, Gengyan Zhao, Tyler Bradshaw, and Alan B. McMillan. A deep learning approach for 18F-FDG PET attenuation correction. *EJNMMI Physics*, 5(1):24, December 2018.

[5] Getao Du, Xu Cao, Jimin Liang, Xueli Chen, and Yonghua Zhan. Medical Image Segmentation based on U-Net: A Review. *Journal of Imaging Science and Technology*, 64(2):020508–1–020508–12, March 2020.

[6] Chengbin Zeng, Yi Liu, and Chunli Song. Swin-CasUNet: Cascaded U-Net with Swin Transformer for Masked Face Restoration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 386–392, Montreal, QC, Canada, August 2022. IEEE.

[7] Ping Li, Shuo Wang, Tang Li, Jingfeng Lu, Yunxin HuangFu, and Dongxue Wang. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis, 2020.