GAN-based synthetic FDG PET images from T1 brain MRI can serve to improve performance of deep unsupervised anomaly detection models

Daria Zotova[1], Julien Jung[2], and Carole Lartizien[1][0000−0001−7594−4231]

[1] Univ Lyon, CNRS, Inserm, INSA Lyon, UCBL, CREATIS, UMR5220, U1206, F-69621, Villeurbanne, France
{Daria.Zotova,Carole.Lartizien}@creatis.insa-lyon.fr
[2] Lyon Neuroscience Research Center, CRNL, INSERM U1028, CNRS UMR5292, University Lyon 1, Lyon, France

Abstract. Research in cross-modal translation or synthesis domain has been very productive over the past few years to tackle the scarce availability of large curated datasets for the training of deep models, with promising performance of GAN-based architectures. However, only a few of these studies assessed task-based related performance of these synthetic data. In this work, we design and compare different GAN-based frameworks for generating synthetic brain FDG-PET images from T1-weighted MRI data, and explore further impact of adding these fake PET data in the training of a deep brain anomaly detection model. Qualitative and quantitative results allow us to conclude that the generated PET images look similar to real ones with SSIM and PSNR values around 0.88 and 23.5 respectively for the best GAN architecture. Training of the brain anomaly detection model on hybrid datasets including 35 real and 40 synthetic FDG PET data, allows achieving a 65% detection sensitivity of subtle epilepsy lesions in 17 real PET exams of patients, while the sensitivity is 53% when training with the 35 real PET exams only, thus demonstrating the diagnostic value of these synthetic data for the design of CAD models.

Keywords: Medical Image synthesis · Unsupervised Learning · CycleGAN · PET MRI · lesion detection

1 Introduction

One major limitation to the performance of deep learning models for medical image analysis is the scarce availability of large annotated training databases. It is all the more difficult to acquire large datasets of paired multi-modality exams (we indeed often have to deal with missing or incomplete datasets) and to sample the variability of the normal and pathological pattern distributions. For this reason, unsupervised or weakly supervised paradigms have gained significant interest over the past few years, due the constraint release on the annotation process [11]. This includes anomaly detection models, which were shown to perform well, especially for detection and segmentation tasks in neuroimaging [3].

This subgroup of methods which consists in learning normal representations or patterns extracted from normal (i.e. non pathological) populations only, allow relaxing the pressure on the annotation of the pathological cases. However, gathering large datasets of normative populations is another challenge, since the vast majority of available clinical images are patient data with pathological patterns. Data augmentation techniques have been proposed as a way to address this issue in the data space [9, 11, 16]. One approach of data augmentation consists of explicit generation of synthetic instance based on synthesis or simulation methods as reviewed in [4]. Challenges in this domain have been addressed over the past few years through international events such as Sashimi workshops held in conjunction with the MICCAI conference. In this study we focus on synthesis methods based on deep neural architectures. The main trend has focused on the use of architectures based on segmentation networks such as U-Net, recently combined with adversarial branches as in generative adversarial networks (GANs) or Cycle-GAN for the synthesis of fake mono- or multi-modal modalities based on tuplets of paired (i.e. coregistered slices or volume of the same patient) or unpaired (i.e. paired of input and output training data that do not belong to the same patient or are not spatially coregistered) different modalities (e.g. multiple sequences of MRI, CT or PET) [16]. Research in this cross-modal translation or synthesis domain has been very productive over the past few years [2, 10, 13–15, 8]. As far as we know, performance of most of the proposed architectures was evaluated based on visual quality metrics only, such as PSNR or turing test, but, only a few assessed task-based related performance, especially in the unsupervised anomaly detection context [15].
The objective of this study is to build on the recent MRI to PET GAN models to design an efficient architecture for the synthesis of realistic PET images derived from T1 MRI images of normal subjects. We do not only evaluate the visual quality of the synthetic PET data with standard metrics such as PSNR but also quantify their added value for the final medical task at hand. To that purpose, we consider the diagnostic task of epilepsy lesion detection in [^{18}F]fluorodeoxyglucose (FDG) PET exams. Following the idea proposed in [1], we build an automated unsupervised anomaly detection model that combines a feature extraction step based on a siamese network and a one-class SVM model. This model is trained on FDG PET brain exams of normal subjects. Our hypothesis is that increasing the size of the training dataset with synthetic FDG exams should translate into a gain in the epilepsy lesion detection rate.The main contributions of this paper are:

- A model derived from the Cycle-GAN architecture for the synthesis of realistic FDG PET exams of normal subjects from T1 MRI.
- A comparison of different variants of GANs methods based on the same training dataset.
- A global evaluation of the quality of these synthetic data including quantitative metrics of visual image quality as well as their ability to mimic real data for a diagnostic lesion detection task at hand.

2 Method

2.1 Synthesis of realistic PET data with GANs

In recent years, generative adversarial networks (GANs) [5] have demonstrated impressive results in computer vision and in biomedical image analysis, for sample generation, image synthesis, quality enhancement and image segmentation [14]. The basic structure of a GAN consists of the generator that is trained to generate new synthetic samples, and the discriminator that tries to distinguish examples being real or fake. These two models are trained simultaneously and compete against each other. In this study, we build on a comparative analysis of different variants of GAN architectures to design the optimal configuration with adapted loss terms for missing PET data generation from T1 MRI data.

GANs architectures

Simple-GAN. We first propose to use a standard GAN architecture with one generator G_B and one discriminator D_B (the upper part in Figure 1). Generator G_B attempts to improve the quality of the translated output x_b of domain B from the original input y_A from the original domain A, thus deceiving the discriminator D_B. The training procedure is formulated as a min-max optimization problem of an objective function that the discriminator is trying to maximise and the generator is trying to minimize. In this study, we implement the least squares GAN (LSGAN) model [7] that aims to minimize the following discriminator $L_{LSGAN}(D_B, A, B)$ and generator $L_{LSGAN}(G_B, A, B)$ losses :

$$\begin{aligned} L_{LSGAN}(D_B, A, B) &= E_{p(x_b)}[D_B(x_b)^2] + E_{p(y_b)}[(D_B(y_b) - 1)^2] \\ L_{LSGAN}(G_B, A, B) &= E_{p(x_b)}[(D_B(x_b) - 1)^2] \end{aligned} \tag{1}$$

where y_a and y_b are true images of domain A and B, respectively, and $x_b = G_B(y_a)$ is the fake image of domain B generated from y_a.

In the context of supervised image translation, where the model can be trained on paired images in both domains at the pixel level (e. g. corresponding images of the same patient), we propose to add a mean squared error (MSE) loss term L_{mse} (see eq. (2)) between the fake image x_b generated from a true image y_a of domain A and its paired true image y_b in domain B.

$$L_{mse}(G_B) = E_{p(x_b)}[(x_b - y_b)^2] \tag{2}$$

Cycle-GAN. Cycle-GAN consists of two generator networks G_A and G_B and two discriminator networks D_A and D_B. The baseline Cycle-GAN model is shown in Figure 1. The generators translate images from domain A to domain B and vice versa. Each of the generator networks is trained adversarially using a corresponding discriminator D_A and D_B. In addition to the adversarial loss term of the simple GAN network in eq. (1), the key element in training Cycle-GAN network is a cycle-consistency loss function L_{cyc}:

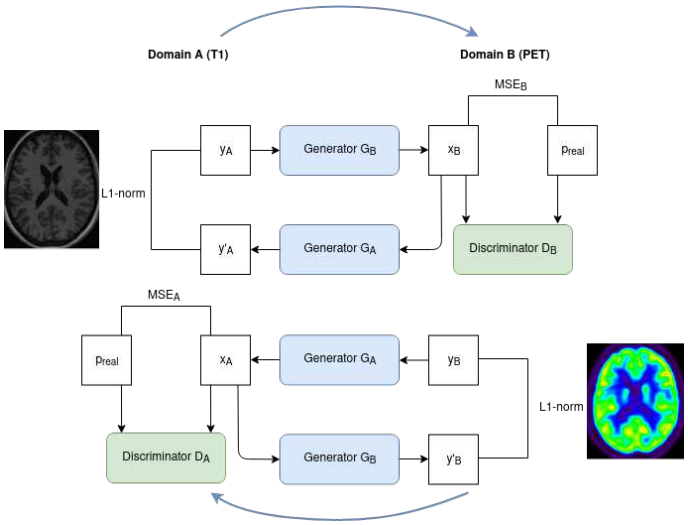

Fig. 1. Cycle-GAN architecture based on two baseline GANs translating images from domain A to domain B (upper GAN) and vice versa (lower GAN).

$$L_{cyc}(G_A, G_B) = E_{p(y_a)}[\| y'_a - y_a \|_1] + E_{p(y_b)}[\| y'_b - y_b \|_1] \quad (3)$$

where y'_a is the fake image of domain A generated by generator G_A from the fake x_b, that is $y'_a = G_A(x_b)$ with $x_b = G_B(y_a)$. As for the simple GAN formulation, in a paired mode, we add a MSE loss term between real and synthetic images of both domains A and B.

Implementation details

We consider two configurations depending on the size of the input data:

- ***semi-3D*** models which receive three adjacent transverse slices as input (each slice corresponding to one channel)
- ***3D-patch*** models where we feed 3D mini patches extracted from the original 3D images into the network

We take ResNet as the backbone architecture of both generators with 9 residual blocks for the semi-3D approach and 2 blocks for the fully 3D-patch configuration. PatchGAN is selected for the discriminators following the architectures proposed in [18]. In the semi-3D configuration, the whole 3D image is reconstructed by stacking the generated transverse slices. In the 3D-patch setting, we crop the generated 3D patches so as to consider only their central part as it has been shown in [6] that predictions for edge pixels have lower accuracy, thus we consider only areas with higher prediction confidence. All patches are then stacked to reconstruct the 3D volume. For both semi-3D and 3D-patch configurations, we finally apply Gaussian smoothing as a post-processing to tackle with "border" effect that may occur when stacking either slices or mini-volumes. All

models were written by using PyTorch version 1.3.1 and we took python code provided by [18] as a baseline.

2.2 Application to the training of a deep epilepsy lesion detection model

We build an epilepsy lesion detection model based on FDG PET exams following the idea from [1]. This model couples efficient patch-based representation learning based on a siamese autoencoder architecture and a OC-SVM anomaly detection algorithm. It is trained on an healthy control population. When tested on an epilepsy patient, it allows outlining the locations of abnormalities with regards to the normal brain population, thus producing anomaly score maps. In this work, we reproduce the 2D siamese architecture depicted in Fig 5 of [1] with 15x15 patch size, resulting in a feature vector of dimension 64 and perform a 4-fold cross validation to extract the latent variables of the control population. Images are scaled between 0 and 1 at image level before feeding the patches to the deep autoencoder architecture. We then build one oc-SVM model per voxel with RBF kernel in the latent representation space learned by the siamese autoencoder.

3 Experiments and results

3.1 Data

This study including three clinical image databases was approved by our institutional review board with approval numbers 2012-A00516-37 and 2014-019 B and a written consent was obtained for all participants. The first database is used to compare the different deep generative models of synthetic PET data. It consists of a series of 35 paired FDG PET and T1 weighted MRI scans co-registered to the MNI space, thus leading to 3D image volumes of size 157x189x136 with $1mm^3$ isotropic voxel size. These data were acquired on 35 healthy volunteers on a 1.5T Sonata scanner and mCT PET scanner (Siemens Healthcare, Erlangen, Germany). The second database consists of 40 T1-weighted MRI exams acquired on healthy control subjects on the same 1.5T MR Sonata scanner. It is used to generate synthetic FDG PET data that then serve to train the brain anomaly detection model introduced in section 2.2. The third database consists of 17 paired FDG PET and T1 weighted MRI scans of patients with confirmed medically intractable and subtle epileptogenic lesions, as illustrated in Figure 3. These data were acquired on on the same 1.5T Sonata MR scanner. All control and patient PET scans are rigidly aligned to their corresponding MRI then co-registered to the MNI space with SPM12 (https://www.fil.ion.ucl.ac.uk/spm/doc/manual.pdf). All original MRI and PET 3D data are scaled between 0 and 1.

3.2 Generation of synthetic PET images from T1 healthy controls

For the semi-3D approach, we explore in total 4 variants of GANs for paired examples based on Simple-GAN or traditional Cycle-GAN architectures both

with and without MSE loss. Forty-six triplets of adjacent slices per patient are extracted thus resulting in around 1 200 training samples for each model. For the 3D-patch based approach, we use Simple-GAN and Cycle-GAN both with additional MSE loss. 6 069 mini-patches of size 32x32x32 are extracted for each healthy subject with a stride of 8, thus leading to more than 200 000 training mini-volumes. A 4-fold cross-validation performance study is conducted with 26 controls in the train set and 9 controls in the validation set. During the training, Structural Similarity Index (SSIM) [12] between real and synthetic validation images serves as a quality metric to define the optimal configuration (early stopping criterion). All semi-3D approaches and 3D-patch models are trained for a maximum of 200 and 100 epochs with a batch size of 5 and 10, respectively, and Adam optimizer. The learning rate of 0.0002 is kept constant for the 3D-patch models, while for the semi-3D models it is kept constant up to 100 epochs and linearly decayed to zero over the next 100 epochs. A 3D Gaussian smoothing is applied on the reconstructed PET images of all model types (semi-3D and 3D-patch) to reduce border effects. Among a range of values between 0 and 3 mm FHWM, the value of 1.5 mm is shown to produce the best SSIM values.
Table 1 reports the mean SSIM, Peak Signal to Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS)[17] with corresponding standard deviation computed over all validation samples and all folds for each of the six considered models. Semi-3D Cycle-GAN with MSE loss is shown to perform the best among the 4 semi-3D models considered in this study. Two-tailed Wilcoxon signed rank tests yield significant differences between the semi-3D and 3D-patch Cycle-GAN models with MSE loss for the PSNR (p-value $< 10^{-6}$) and LPIPS (p-value $< 2\text{x}10^{-4}$) metrics. A p-value of 0.069 is also achieved for the SSIM metric. Also note that our proposition to add the MSE loss term to the Cycle-GAN global loss allows a significant improvement of all three metrics.

Table 1. Average visual quality metrics computed on the 35 synthetic PET exams generated from T1 MRI of 35 healthy subjects.

Configuration	**Model**	**SSIM**	**PSNR**	**LPIPS**
semi-3D	Simple-GAN	0.818±0.021	19.655±1.441	0.035±0.008
	Simple-GAN with MSE loss	0.879±0.021	23.177±1.760	0.022±0.005
	Cycle-GAN	0.837±0.028	21.750±1.142	0.030±0.006
	Cycle-GAN with MSE loss	**0.883±0.022**	**23.525±1.388**	**0.022±0.005**
3D-patch	Simple-GAN with MSE loss	0.869±0.031	19.852±0.597	0.034±0.017
	Cycle-GAN with MSE loss	0.875±0.023	17.760±1.767	0.026±0.007

In the following, we consider the best performing of each configuration, namely semi-3D and 3D-patch Cycle-GAN models with MSE loss. Example synthetic PET data generated by these two configurations of Cycle-GAN models from the same T1 MRI of a control subject are illustrated in Figure 2 and compared with the reference PET image of this subject. Both models allow generating visually realistic FDG PET data.

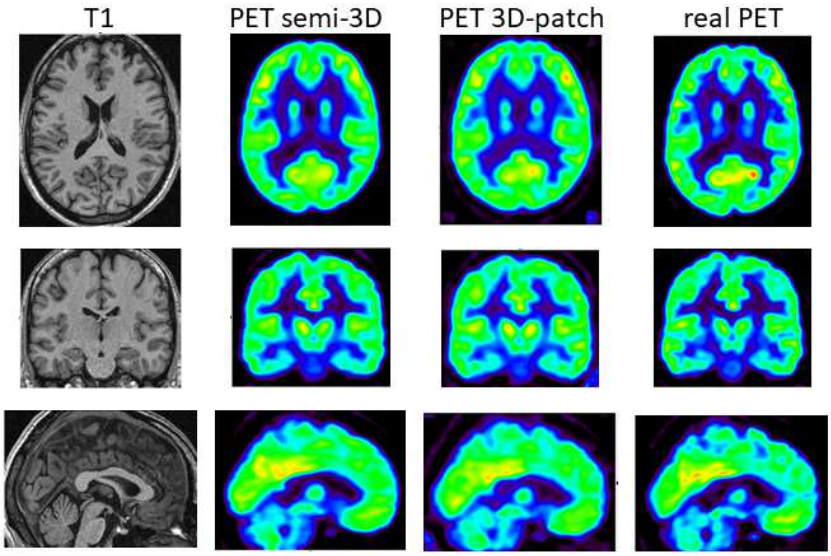

Fig. 2. Qualitative result on one control subject. Left column: original T1 MRI, right column: original PET image, the two central columns from left to right: synthetic PET image generated with semi-3D and 3D-patch Cycle-GAN models with MSE loss.

3.3 Application of synthetic PET data to the training of a brain anomaly detection model for epilepsy patients screening

Our second objective is to demonstrate that the realistic synthetic PET data can serve to improve performance of machine learning based diagnostic models. The considered application is the brain anomaly detection model described in section 2.2. This model is trained on three different databases: the series of 35 real control PET dataset described in section 3.1 and two hybrid databases mixing these 35 real control PET with 40 synthetic control FDG PET data generated by the semi-3D and 3D-patch Cycle-GAN models (with MSE loss), respectively. These 3 models are then tested on 17 patients with confirmed medically intractable epileptogenic lesions. Note that these patients correspond to difficult detection cases. Their FDG PET exam is indeed considered as normal, meaning that the hypometabolic lesions are subtle and barely visible by naked eye. Results reported in Figure 4 indicate that the best detection sensitivity of 64.7% was achieved with the model trained on the hybrid dataset including the 40 PET data generated from the best semi-3D model. Adding these synthetic data to the training, which here amounts to doubling the number of training samples, allows a 20% gain in sensitivity compared with that achieved with the same model trained on 35 real PET scans only achieving a 53% sensitivity. Performance achieved with the hybrid dataset including 40 PET data generated from the 3D-patch model reached a 41% sensitivity which is lower than that achieved with the model trained on the 35 real PET exams. Figure 3 illustrates anomaly maps derived from the three detection models on three test epilepsy patients.

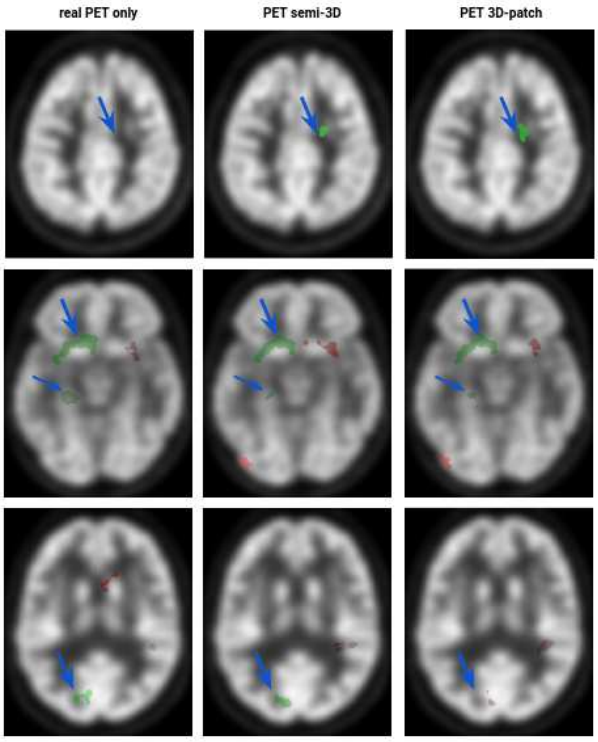

Fig. 3. Example cluster maps for three patients produced by the detection models, from left to right: 35 real PET scans, 35 real+40 synthetic PET (semi-3D Cycle-GAN), 35 real+40 synthetic PET (3D-patch Cycle-GAN). Blue arrows point to suspicious anatomical regions. The upper line demonstrates a case where both models trained with additional synthetic data managed to detect a lesion with a high confidence (bright green color) in the right internal frontal lobe, while it is missed by the model trained on real PET data solely. The middle line shows a successful case, where all models detect two clusters (green coloured) in the left internal temporal lobe and hippocampus. For the bottom line patient, models trained on original PET and with synthetic PET (semi-3D) managed to detect a lesion in the left lateral remainder of occiptal lobe, but the correct location is missed for the 3D-patch Cycle-GAN model. Red clusters correspond to false positives (the brighter the color the higher the rank of the cluster).

4 Discussion and conclusion

In this study, we demonstrate that realistic FDG PET exams of healthy subjects can be generated from GAN based architectures with T1 MRI as input. We also show that these synthetic data could efficiently serve as training samples to boost the performance of machine learning based diagnostic models.

As seen in Figure 2, both semi-3D and 3D-patch Cycle-GAN models produce PET images which closely match the original ones. In both cases, however, the histogram of the intensity does not perfectly match that of the original data. One perspective regarding this issue would be to further constrain the

generative model to match the global intensity value of the true PET image or perform histogram matching. This may positively impact the performance of the brain anomaly detection model. Figure 4 shows that performance achieved with the added synthetic 3D-patch data are lower than that achieved with the true PET training dataset. Paired visual analysis of the anomaly maps generated by models trained with semi-3D and 3D-patch synthetic PET indicate very similar patterns for all 17 patients, except for two of them which did not contain any suspicious cluster in the lesion anatomical region for 3D-patch unlike in the produced anomaly map based on the semi-3D approach. Further analysis is required to better understand this observed difference. We also plan to add more patient to this analysis to evaluate if the trend observed in Figure 4 is confirmed.

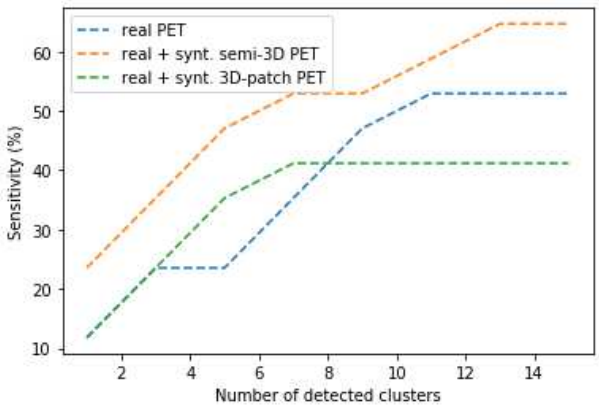

Fig. 4. Comparative detection curves estimated on the 17 patients of the test dataset based on the brain anomaly detection models trained on the three considered databases. x-axis: number of detected clusters per patient based on individual thresholding of the score maps outputted by the detection model, y-axis: sensitivity.

The best performing model allows achieving a detection sensitivity of 64% which may seem low. Note that this value has to be compared with very low sensitivity of human experts on these difficult diagnostic cases and is in par with reported values in a recent study questioning the added value of synthetic PET data for the same clinical application [15].

References

1. Alaverdyan, Z., Jung, J., Bouet, R., Lartizien, C.: Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. Medical Image Analysis **60**, 101618 (2020)
2. Armanious, K., Jiang, C., Abdulatif, S., Küstner, T., Gatidis, S., Yang, B.: Unsupervised medical image translation using cycle-medgan. In: 2019 27th European Signal Processing Conference (EUSIPCO). pp. 1–5. IEEE (2019)

3. Baur, C., Denner, S., Wiestler, B., Albarqouni, S., Navab, N.: Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. arXiv:2004.03271 [cs, eess] (Apr 2020)
4. Frangi, A.F., Tsaftaris, S.A., Prince, J.L.: Simulation and synthesis in medical imaging. IEEE transactions on medical imaging **37**(3), 673–679 (2018)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
6. Huang, B., Reichman, D., Collins, L.M., Bradbury, K., Malof, J.M.: Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations. arXiv preprint arXiv:1805.12219 (2018)
7. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE conference on computer vision (ICCV). pp. 2813–2821 (2017)
8. Pan, Y., Liu, M., Lian, C., Zhou, T., Xia, Y., Shen, D.: Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's disease diagnosis. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 455–463. Springer International Publishing, Cham (2018)
9. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 60 (2019)
10. Sikka, A., Peri, S.V., Bathula, D.R.: Mri to fdg-pet: cross-modal synthesis using 3d u-net for multi-modal alzheimer's classification. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 80–89. Springer (2018)
11. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020)
12. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
13. Wei, W., Poirion, E., Bodini, B., Durrleman, S., Ayache, N., Stankoff, B., Colliot, O.: Predicting pet-derived demyelination from multimodal mri using sketcher-refiner adversarial training for multiple sclerosis. Medical image analysis **58**, 101546 (2019)
14. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Generative adversarial networks for noise reduction in low-dose ct. IEEE transactions on medical imaging **36**(12), 2536–2545 (2017)
15. Yaakub, S.N., McGinnity, C.J., Clough, J.R., Kerfoot, E., Girard, N., Guedj, E., Hammers, A.: Pseudo-normal pet synthesis with generative adversarial networks for localising hypometabolism in epilepsies. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 42–51. Springer (2019)
16. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis **58**, 101552 (2019)
17. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
18. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)