# UCoAtCycleGAN : UNet based CoAtNet Cycle GAN

Zhaoyang Li[1]
University of Wisconsin-Madison
zli2344@wisc.edu

Boqi Zhao[1]
University of Wisconsin-Madison
bzhao78@wisc.edu

Yingwei Song[1]
University of Wisconsin-Madison
ysong279@wisc.edu

Yiyang Wang[1]
University of Wisconsin-Madison
wang2687@wisc.edu

## Abstract

*The breakthrough already achieved by the generation model in medical imaging shows its potential to be used for translation between different imaging modalities. To reduce the financial burden on the patient and to speed up the diagnostic process, we propose a novel unsupervised GAN framework to achieve the goal of using magnetic resonance imaging (MRI), an inexpensive and easy-to-use technique, to generate certain types of images of positron emission tomography (PET), which is an expensive and complex time-consuming technique. Our key idea is to use CycleGAN to map the input image to a specified one in the output domain and use a combination of MobileNetV2 and self-attention in the process to achieve global and local feature capture of the input photo. We present experiments on the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset with 726 training and 80 testing subjects and obtain acceptable performance in PET image synthesis. We also use various metrics to evaluate the generated images. If you want to learn more about our project, please refer to* `https://github.com/GeofrreyLi/U-TransCyGan`

## 1. Introduction

Compared with CT (Computerized tomography), Positron Emission Tomography (PET) contributes more to helping doctors to diagnose some neurological diseases. But not only the fees for patients to take a PET are a considerable burden, and the radio-tracer injection will make the process more complicated. Therefore, We are wondering if we can use another cheaper and relatively easy-to-operate technique. Transferring from Magnetic Resonance Imaging (MR) to PET is a good practice [20]. Because MR is a relatively cheap method and doesn't cause harm as strongly as PET to brains and MR will be the first step in MR/PET system [4]. In this paper, we will introduce the model we have built to make this transformation.

Since Generative adversarial networks (GANs) [6] has earned a great influence in the image synthesis area. And the GANs model has been widely used in the processing of medical images. But in the process of medical image processing, we will encounter many very difficult problems. For example, the amount of data in general medical images is very small [14]; how to obtain more information from the limited training data? In this article, we used data enhancement: by adding different forms of noise to the data set to increase the data set to achieve better results. And our model is not just a one-way transition from MR to PET, and our model is a two-way transition. Compared with a one-way transition from MR to PET, we can double the available data set and enhance this tool's flexibility. Whether the doctor wants to get PET through MR or get MR from PET, we can get corresponding support in our model.

Another difficulty is that for medical image data, the large amount of information appearing in single data is also a big problem. In particular, a PET image is a three-dimensional image of the brain. This leads to an excessive amount of information, and it is difficult to extract more important features. If Convolution is used, it will cause the model to focus too much on local features, and if relative self-attention is used, it will cause the model to focus too much on global features and increase the amount of calculation. So we combined convolution, and relative self-attention [18] into our model. This makes our model not only have the ability to extract global features but also the ability to extract local features. And the convolution layer we use is the convolution layer in MobileNetV2 [3], which can further reduce the computational load of our model. Make the model more lightweight.

In order to enable the mutual conversion between MR and PET images, in this article, we draw on the architecture of CycleGAN [2]. But we used MobileNetV2 and relative self-attention as the main units, and for the discriminator,

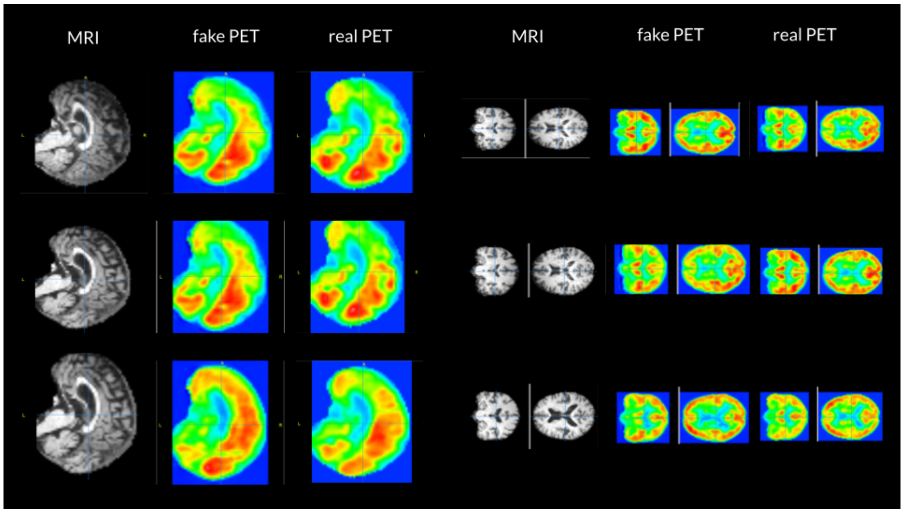

Figure 1. The Generated Image

we did not simply use a patchGAN but connected the half of the input information captured by CycleGAN to an FCN Layer as the output judgment value for the discriminator to judge. For Generator, our MR is equivalent to input as content information so that the model can generate corresponding PET pictures. And CycleGAN can allow our models to be converted to each other. We make the content information parameters learned in the first half of the model appear in the second half of the generator, which improves the learning efficiency of the model.

### 1.1. Main Contributions

Our main contributions are: 1. Introduce the GANs model to the conversion of MR and PET images, and allow the two images to be converted to each other. Compared to previous research, our image performance and metrics have improved significantly. 2. Incorporating MobileNetV2 and Relative Attention, the model can capture not only global features but also local features and make the model more lightweight. 3. Using the architecture of CycleGAN, MR and PET can be converted to each other, thereby expanding the amount of data that can be trained. For the Discriminator under the CycleGAN architecture, FCN is used as the last output layer to make the result more stable.

## 2. Related Work

### 2.1. CoAtNet

#### 2.1.1 MobileNetV2

Convolutional Networks have dominated neural architectures for many computer vision tasks, including classification and object detection. Depthwise convolutions [3] are popular in mobile platforms due to the optimization of computational cost and the reduction of parameter size. MobileNetV2 [17], built on depthwise convolutions, is based on an inverted residual structure where the residual connections are between the bottleneck layers. The bottlenecks encode the model's intermediate inputs and outputs, while the inner layer encapsulates the model's ability to transform from lower-level concepts, such as pixels, to higher-level descriptors. In addition, as with traditional residual connections, shortcuts enable faster training and better accuracy. Moreover, another advantage of Mobile convolution is that it has a strong connection with Transformer blocks. So we mostly use Mobile convolutions in this paper.

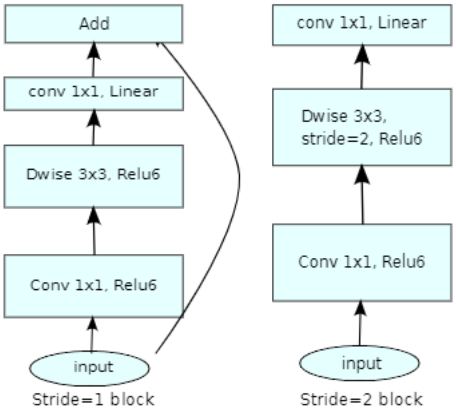

Figure 2. Mobilenet Architecture [17]

#### 2.1.2 Self-Attention with Relative Position Representations

With the attention mechanism, the Transformer can work well for different tasks. However, Transformer does not explicitly model relative or absolute position information in its structure and needs to add a representation of absolute positions to its input. Self-attention with relative position representations [19] is a mechanism that can efficiently consider representations of relative positions or distances between sequence elements. It incorporates relative position representations into the Transformer self-attentive mechanism, where residual connections help to propagate position information to higher layers.

### 2.2. U-Net

It is a convolutional neural network initially developed for biomedical image segmentation, such as BRATS. The network of U-Net [15] consists of a contracting path (encoder) and an expansive path (decoder). During the contraction, the spatial information is reduced while feature information is increased. As it uses the successive layers with upsampling operators, the resolution of the output is increased by these layers. The expansive pathway combines the feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path. This architecture will be used in our generators of CylceGAN due to its ability to assemble precise output.

#### 2.2.1 Skip Connections and Bridge

Skip connections provide additional information that helps the decoder to generate better semantic features. They also act as a shortcut connection that helps the indirect flow of gradients to the earlier layers without any degradation. The bridge connects the encoder and the decoder network and completes the flow of information.

### 2.3. CycleGAN

CycleGAN [24] is a technique that involves automatic training of image-to-image translation models without paired examples. The model is trained in an unsupervised manner using a collection of images from the source and target domain that do not need to be related in any way. It achieves the goal of learning a mapping G: $X \to Y$ such that the distribution of images from G(X) is indistinguishable from the distribution of Y using an adversarial loss, coupling it with an inverse mapping F: $Y \to X$ and using a cycle consistency loss to push $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. In this architecture, there are two GAN models. In this paper, the input of the generator from GAN1 is the collection of MRI images, and the output of it is the generated PET images. The input of the discriminator from GAN1 is the PET images from the collection of PET images and the generated PET images. The output is the likelihood of judging if the image generated is real (from the collection of PET images). Inversely, the architecture of GAN2 is the same as GAN1, with all input and output reversed.

### 2.4. WGAN-GP

Wasserstein GAN + Gradient Penalty [8] is a GAN model that uses the Wasserstein loss formulation plus a gradient norm penalty to achieve Lipschitz continuity. As the original WGAN uses weight clipping to achieve 1-Lipschitiz functions, it will lead to undesirable behavior by creating pathological value surfaces and capacity underuse and gradient explosion or vanishing under some circumstances. In this case, WGAN-GP replaces weight clipping with gradient penalty as a soft version of the Lipschitz constraint. Gradient penalty works because a differentiable function f is 1-Lipschitz if and only if it has gradients with the norm at most one everywhere. In addition, WGAN-GP can create a complex boundary to surround the modes of the model, which has a better performance than WGAN in our task.

## 3. Method

The overall architecture for our method is based on the CycleGAN, which aims to implement two translation works: from PET image to MR image, and then translate the MR image back to PET image. Each individual GAN consists of a generator and a discriminator. The generator used the U-Net architecture and CoAtNet, while the discriminator used the combination of CoAtNet and FCN architecture.

### 3.1. U-Net Based generator

The promising results in previous medical vision work, especially for medical image segmentation tasks, gave us the idea of applying the U-Net architecture to the generator in a GAN. U-Net consists of convolution operation, max Pooling, ReLU Activation, concatenation, up-sampling layers, and three sections: contraction, bottleneck, and expansion. This is very similar to an encoder-decoder architecture in image generative models. In other words, U-Net is similar to an auto-encoder that learns latent representations and reconstructs the output with the same size as the input [1]. The generator's structure is based on the UNET, but we introduced convolution and self-attention (CoAtNet) to replace the original UNET backbone. This approach of marring convolution and attention, using the C-C-T-T, has been demonstrated to save computation and memory usage [9]. substantially. The skip connection in UNET is one of the core ideas that help preserve the information from layer to layer. Low-level information can be avoided from being washed away by sequences of convolutions and passed to the output layer.

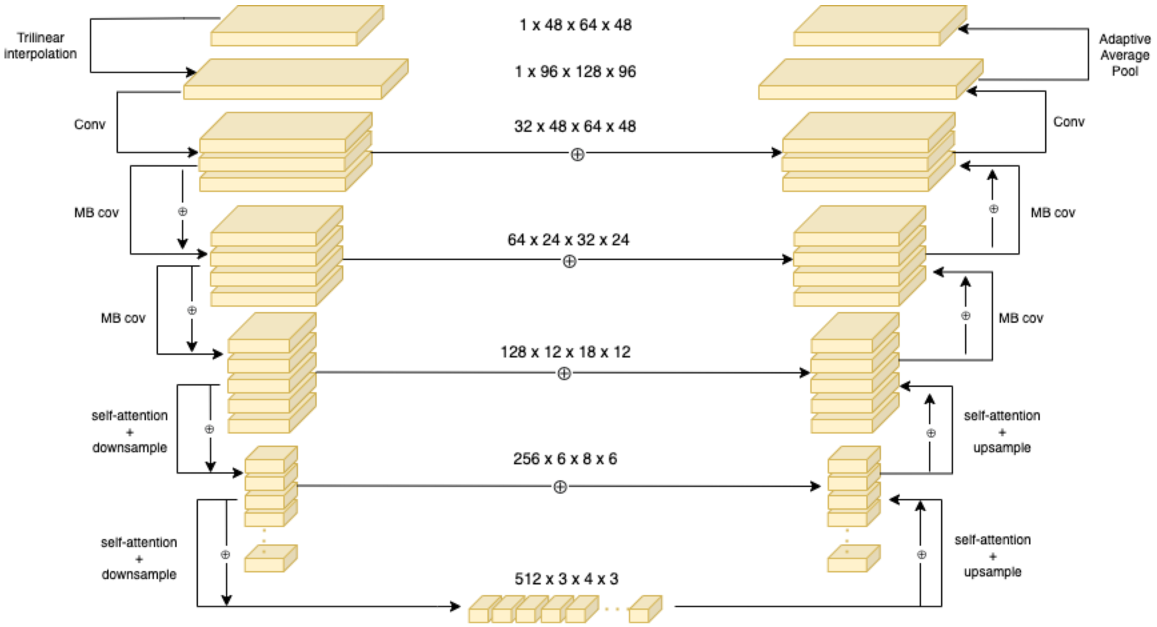

Figure 3. Generator Architecture

## 3.2. Discriminator

The discriminators we used are not actually the conventional PatchGANs. In brief, we used the CoAtNet+FCN architecture here to discriminate whether the generated images are true/false.

Actually, our discriminator is a brand new model based on the left half of our Generator and uses a fully connected layer at the end to generate a signal value to judge the authenticity of the image.

For the discriminator architecture, after entering an input, we use Trilinear Interpolation to increase the size of a single data. This step can increase the robustness of the model and provide Data Augmentation to a certain extent because each incoming data will be inserted into different Pixels, so even the same input will produce different pictures after Interpolation. After that, we use convolution to increase its dimension. After that, we used two MobileNetV2 Convolutions [1]. This technology uses a residual method to reduce computational complexity (for details, please refer to our related work). Next, we used convolution+max pooling to downsample data, and then used self-attention [3] to extract global information.

All in all, our discriminator can not only enhance the robustness of the model but also consider the global and local feature characteristics, which is a huge improvement for our model.

## 3.3. Loss

$$\mathrm{L}_{identity}(G,F) = \mathbb{E}_{y\sim p_{data}(y)}[||G(y) - y||_1] + \mathbb{E}_{x\sim p_{data}(x)}[||F(x) - x||_1]$$

$$\mathcal{L}_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}[||F(G(x)) - x||_1] + \mathbb{E}_{y\sim p_{data}(y)}[||G(F(y)) - y||_1]$$

$$\mathcal{L}_{adv} = \underbrace{\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)]}_{\text{Original critic loss}} + \underbrace{\lambda_{gp}\mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[||\nabla_{\hat{x}D(\hat{x})||_2-1)^2}]}_{\text{gradient penalty}}$$

Since the overall architecture is based on the CycleGAN, there are two losses are necessary for training an original CycleGAN: an adversarial loss and a cycle consistency loss, while both are necessary to guarantee good results. For adversarial loss, we used the WGAN-GP loss function, which is an improved loss function and has proven good effectiveness in avoiding the failure of coverage. A cycle consistency loss is also added to ensure that the generated image domain can be changed while the content remains. After that, we introduced an identity loss function to help to preserve color composition in the image-to-image generation.

### 3.3.1 Adversarial loss

Adversarial loss is another core idea in the GAN model, which why makes the GAN is called the “adversarial” network. For the mapping function $G : X \rightarrow Y$ and its discriminator $D_Y$, the objective is trying to find an objective as:

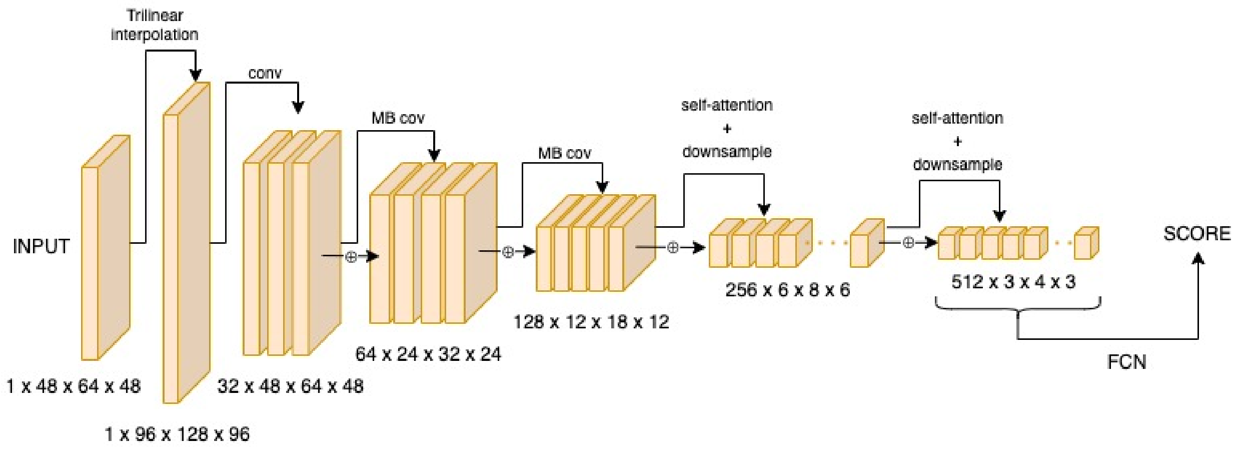

Figure 4. Discriminator Architecture

$$\mathrm{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_y \sim p_{data}(y)[logD_Y(y)] + \mathbb{E}_x \sim p_{data}(x)[log1 - D_Y(G(x))]$$

Where the ideal G is to generate images G(x) that from domain Y, and the $D_Y$ is aimed to discriminate it from the real image $y$. While G aims to minimize the discriminator's loss, D tries to maximize it [24].

In the conventional GAN model, there are cases where it generates poor samples or fails to converge. In the previous work, WGAN was introduced to solve this problem, and they used a weight clipping method to guarantee the Lipschitz continuity on the critic. However, such a method limits the discriminator's weight to a set interval to ensure that the discriminator weights do not change with large oscillations. In later research, the case failing to converge may still happen with this approach. Thus, WGAN-GP was introduced to constrain the gradient norm of the critic. In other words, the purpose of using the gradient penalty is to keep the gradient at a reasonably small value to avoid the discriminator from not converging with large oscillations during training.

$$\mathcal{L}_{adv} = \underbrace{\mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r}[D(x)]}_{\text{Original critic loss}} + \underbrace{\lambda_{gp} \mathop{\mathbb{E}}_{\hat{x} \sim \mathbb{P}_{\hat{x}}}[||\nabla_{\hat{x} D(\hat{x})||_2 - 1)^2}]}_{\text{gradient penalty}}$$

Where $\tilde{x} \sim \mathbb{P}_r$ is the data distribution and $\tilde{x} \sim \mathbb{P}_r$ is the model distribution implicitly defined by $\tilde{x} = G(z), z(z)$, while the z is from some simple noise distribution p [7].

To optimize the GAN and WGAN, Gulrajani et al. introduced a method that enforces a soft version of the constraint with a penalty on the gradient norm for random samples $\hat{x} \sim \mathbb{P}_{\hat{x}}$ [7]. They implicitly defined $\mathbb{P}_{\hat{x}}$ sampling uniformly along the straight lines between pairs of points sampled from the $\tilde{x} \sim \mathbb{P}_r$, which is the data distribution, and the generator distribution $\tilde{x} \sim \mathbb{P}_g$. While enforcing the unit gradient norm constraint everywhere is difficult to achieve, they showed that enforcing it only along these straight lines seems sufficient for a good performance. $\lambda_{gp}$ here is the penalty coefficient; we use $\lambda_{gp} = 10$ for here for the reason that it has been proven to work well across various CNN architectures [7].

### 3.3.2 Cycle consistency loss

$$\mathrm{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + \mathbb{E}_{y \sim p_{data}(y)}[||G(F(y)) - y||_1]$$

The prompt idea here is why we need a cycle consistency loss here is we are actually doing an image-to-image translation. If we can translate PET images into an MR images, then we should also translate the MR image back to a PET images. The role of the cycle consistency loss is to guarantee that the generated image is actually a version, or style, of the input image where only the domain changes, but the semantic contents of the images can be kept. Without a cycle consistency loss, the network can map the same set of input images to any random permutation of images in the target domain [24].

### 3.3.3 Identity loss

$$\mathrm{L}_{identity}(G, F) = \mathbb{E}_{y \sim p_{data}(y)}[||G(y) - y||_1] + \mathbb{E}_{x \sim p_{data}(x)}[||F(x) - x||_1]$$

In the previous work, Zhu et al. [24] proposed introducing an identity loss function, which can encourage color composition to be preserved between input and output. The specific method is to regularize the generator to be near, and identity mapping when the real samples of the target domain are provided as the input to the generator [24]. Without the identity loss function, the generated image's color information will not be constrained and free to change, which is apparently not what we want. For our PET  MR image generation task, though MR images are greyscaled, it has 256 gray level values, and such color composition can be present as the brightness. The main reason why introducing identity loss here is to prevent reverse color in the result.

# 4. Experiments

We use the ADNI dataset to train and evaluate the UCoAtCycleGAN model. We compare with the existing model including, cGAN [13], UcGAN [16], C-VAE [5], and DUAL-GLOW [20]. The performance of our model is fine and meaningful for the clinical situation, and the cycle generation is flexible for real clinical situations. c Our model can potentially be used in Alzheimer's Disease diagnosis.

## 4.1. Dataset

Driven from DUAL-GLOW [20], we use The Alzheimer's Disease Neuroimaging Initiative (ADNI), which provides a large amount of data available for scholars to conduct research on Alzheimer's disease. The ANDI dataset contains many types of medical brain scan images, including cognitively normal (CN), significant memory concern (SMC), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI) or having Alzheimer's Disease (AD). We can obtain correctly matched pairs of FDG-PET and T1-weighted MRIs from the ANDI database.

## 4.2. Data Preprocessing

Because the original MR and PET images contain many values, they are unsuitable for direct training. We use SPM12 for pre-processing. The MR images were nonlinearly mapped to the MNI152 template. Finally, the PET images were mapped to the standard MNI space using the same forward warping identified in the MR segmentation step. The voxel size was fixed at 1.5 × 1.5 × 1.5 mm3 for all volumes, and the final volume size was 64 × 96 × 64 for both MR and PET images.

## 4.3. Data Augmentation

Due to the design of Self-attention, after the calculation of Self-attention. The dimensionality of the image does not increase, i.e., the dimensionality of the features does not increase, so a large amount of data is required. To remove this obstacle, data augmentation methods [21] are introduced to guide us in training UCoAtCycleGAN in a data-efficient way.

Traditional CNN-based GAN models rarely use data augmentation. However, some data augmentation methods based on GAN models [11, 23] have been proposed to train small-order magnitude data samples. Inspired by paper [10], we also use differential augmentation (DiffAug) [12] with three basic operators Translation, Cutout, Interpolation to improve performance.

## 4.4. Implementation

we follow the previously mentioned methodology. For the loss function, we combine the identity loss [25], cycle consistency loss [25], and WGAN-GP loss [8]. And for hyperparameters of our model, we adopt $\lambda_{indentity} = 0.5$, $\lambda_{cycle} = 10$, and $\lambda_{penalty} = 10$, which are the hyperparameters of identity loss, cycle consistency loss, and gradient penalty respectively. we train UCoAtCycleGAN using A100 GPUs and adopt a learning rate of $2e - 4$ for both the generator and discriminator. We use Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.99$. We use Cosine annealing, which is the learning rate decay schedule. We follow the setting to use the ANDI dataset with 726 images to train and 80 images to test. We choose DiffAug [12] as the basic data augmentation strategy during the training process. We use commonly used quantitative indicators for evaluation steps to calculate and compare the test data. To be more specific, we use Mean Absolute Error (MAE), Correlation Coefficients (CorCoef), Peak Signal-to-Noise Ratio (PSNR), and Structure Similarity Index (SSIM) to evaluate generated images. For Cor Coef, PSNR and SSIM, higher values indicate better generation of PET images. Table 1 shows the result, and we will analyze these metrics in more detail in the next section, 4.5.

## 4.5. Numeric Evaluation of Generation

Table 1 compares various generative models, including GAN-baed and flow-based models. As a result, we show that UCoAtCycleGAN gets a good score in CorCoef with 0.969 and SSIM with 0.85, but the PSNR is not as well as the models we compare; this indicates that the reconstruction quality of images and videos affected by lossy compression of our generated images still needs to be improved. And we find that DUAL-GLOW [20] is also better than all of the models, and we will try to improve UCoAtCycleGAN to let it get in touch with DUAL-GLOW. And in table 2, which shows the MAE in all models, we show that the MAE of our model is smaller than cGAn and UcGAN but larger than C-VAE, pix2pix, and DUAL-GLOW. With all metrics above, we conclude that our model has room for optimization and improvement, so we will focus on reducing MAE and increasing CorCeof, PSNR, and SSIM values.

| Method | cGAN | UcGAN | C-VAE | pix2pix | DUAL-GLOW | Ours |
|---|---|---|---|---|---|---|
| CorCoef | 0.956 | 0.963 | **0.980** | 0.967 | **0.975** | **0.969** |
| PSNR | 27.37 | 27.84 | 28.69 | 27.545 | **29.56** | 25.02 |
| SSIM | 0.761 | 0.780 | 0.817 | 0.783 | **0.989** | **0.85** |

Table 1. The table of scores of measurements for evaluating generation

| Method | cGAN | UcGAN | C-VAE | pix2pix | DUAL-GLOW | Ours |
|---|---|---|---|---|---|---|
| MAE | 0.020 | 0.018 | 0.013 | 0.014 | 0.012 | 0.015 |

Table 2. The MAE Table

## 5. Conclusions

This paper introduces examples of applying GANs models in the medical field. And in this subdivision of MR to PET image conversion, the model's performance is improved, and not only one-way conversion, but our model also allows the mutual conversion of two images. This is due to CycleGAN combining MobileNetV2 and Relative Attention and our changes to Discriminator. It makes the calculation efficiency of the model higher and provides a solution for medical image processing characterized by small data sets.

### 5.1. Limitations

For this model, there are still some shortcomings. Because the amount of medical image data is very small, even if we use data enhancement technology and CycleGAN's two-way learning, the amount of data is still too small for the normal GAN model training, which may cause the model to be over-distributed for the data set. Likelihood of fit.
For MR and PET images, too large 3D images will greatly challenge information capture and the effective use of computing resources. The large single image size greatly increases our calculation time, and we have to use more layers to capture more orientation information. This directly leads to an increase in the training time of our model and the problem of parameter explosion

### 5.2. Future Work

For future work, we will look for larger data sets and strengthen data enhancement processing so that the model can have more data for training. Because the parameters on the left and right sides of CycleGAN's Generator interact, we also want not simply to stack them, but to allow them to interact so that the right side can be more affected by the input image during the up-sampling process. And we also plan to continue to optimize the model architecture. Because this is a CycleGAN, the architecture of CycleGAN's Generator enables us to convert images to each other. We found that this is very similar to the process of Encoder and Decoder [22]. We plan to add a gap between the original image and the image after two conversions in the penalty item so that our model can have a more objective and effective penalty item and performance indicator.

## References

[1] Xiaocong Chen, Yun Li, Lina Yao, Ehsan Adeli, and Yu Zhang. Generative adversarial u-net for domain-free medical image augmentation, 2021.

[2] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[3] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[4] Alexander Drzezga, Michael Souvatzoglou, Matthias Eiber, Ambros J Beer, Sebastian Fürst, Axel Martinez-Möller, Stephan G Nekolla, Sibylle Ziegler, Carl Ganter, Ernst J Rummeny, et al. First clinical experience with integrated whole-body pet/mr: comparison to pet/ct in patients with oncologic diagnoses. *Journal of Nuclear Medicine*, 53(6):845–855, 2012.

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8857–8866, 2018.

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial networks. corr abs/1406.2661 (2014). *arXiv preprint arXiv:1406.2661*, 2014.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[9] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up, 2021.

[10] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[12] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxin Yang. Differentiable automatic data augmentation. In *European Conference on Computer Vision*, pages 580–595. Springer, 2020.

[13] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[14] Felix Ritter, Tobias Boskamp, André Homeyer, Hendrik Laue, Michael Schwier, Florian Link, and H-O Peitgen. Medical image analysis. *IEEE pulse*, 2(6):60–70, 2011.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[18] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[19] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018.

[20] Haoliang Sun, Ronak Mehta, Hao H Zhou, Zhichun Huang, Sterling C Johnson, Vivek Prabhakaran, and Vikas Singh. Dual-glow: Conditional flow-based generative model for modality transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10611–10620, 2019.

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[22] Kang Xie, Mengting Luo, Hu Chen, Mingming Yang, Yuhua He, Peixi Liao, and Yi Zhang. Speckle denoising of optical coherence tomography image using residual encoder–decoder cyclegan. *Signal, Image and Video Processing*, pages 1–13, 2022.

[23] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.