DUAL-GLOW: Conditional Flow-Based Generative Model for Modality Transfer

Haoliang Sun[1,2,3], Ronak Mehta[1], Hao H. Zhou[1], Zhichun Huang[1],
Sterling C. Johnson[1], Vivek Prabhakaran[1], and Vikas Singh[1]

haolsun.cn@gmail.com, ronakrm@cs.wisc.edu, {hzhou97,zhuang294,prabhakaran}@wisc.edu,
scj@medicine.wisc.edu, vsingh@biostat.wisc.edu
[1]University of Wisconsin-Madison [2]Shandong University [3]Inception Institute of Artificial Intelligence

Abstract

Positron emission tomography (PET) imaging is an imaging modality for diagnosing a number of neurological diseases. In contrast to Magnetic Resonance Imaging (MRI), PET is costly and involves injecting a radioactive substance into the patient. Motivated by developments in modality transfer in vision, we study the generation of certain types of PET images from MRI data. We derive new flow-based generative models which we show perform well in this small sample size regime (much smaller than dataset sizes available in standard vision tasks). Our formulation, DUAL-GLOW, is based on two invertible networks and a relation network that maps the latent spaces to each other. We discuss how given the prior distribution, learning the conditional distribution of PET given the MRI image reduces to obtaining the conditional distribution between the two latent codes w.r.t. the two image types. We also extend our framework to leverage "side" information (or attributes) when available. By controlling the PET generation through "conditioning" on age, our model is also able to capture brain FDG-PET (hypometabolism) changes, as a function of age. We present experiments on the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset with 826 subjects, and obtain good performance in PET image synthesis, qualitatively and quantitatively better than recent works.

1. Introduction

Positron Emission Tomography (PET) images provide a three-dimensional image volume reflecting metabolic activity in the tissues, e.g., brain regions, which is a key imaging modality for a number of diseases (e.g., Dementia, Epilepsy, Head and Neck Cancer). Compared with Magnetic Resonance (MR) imaging, the typical PET imaging procedure usually involves radiotracer injection and a high cost associated with specialized hardware and tools, logistics, and expertise. Due to these factors, Magnetic Resonance (MR) imaging is much more ubiquitous than PET imaging in both clinical and research settings. Clinically, PET imaging is often only considered much further down the pipeline, after information from other non-invasive approaches has been collected. It is not uncommon for many research studies to include MR images for *all* subjects, and acquire specialized PET images only for a *smaller subset* of participants.

Other use cases. Leaving aside the issue of disparity in costs between MR and PET, it is not uncommon to find that due to a variety of reasons other than cost, a (small or large) subset of individuals in a study have *one or more image scans unavailable*. Finding ways to "generate" one type of imaging modality given another is attracting a fair bit of interest in the community and a number of ideas have been presented [34]. Such a strategy, if effective, can increase the sample sizes available for statistical analysis and possibly, even for training downstream learning models for diagnosis.

Related Work. Modality transfer can be thought of "style transfer" [6, 11, 15, 16, 19, 22, 24, 25, 27, 30, 31, 42, 43, 49, 50] in the context of medical images and a number of interesting results in this area have appeared [13, 17, 23, 28, 32, 34, 44]. Existing methods, mostly based on deep learning for modality transfer, can be roughly divided into two categories: Auto-encoders and Generative Adversarial Networks (GANs) [3, 12, 18]. Recall that auto-encoders are composed of two modules, encoder and decoder. The encoder maps the input to a hidden code h, and the decoder maps the hidden code to the output. The model is trained by minimizing the loss in the output Euclidean space with standard norms (ℓ_1, ℓ_2). A U-Net structure, introduced in [36], is typically used for leveraging local and hierarchical information to achieve an accurate reconstruction. Although the structure in auto-encoders is elegant with reasonable efficiency and a number of authors have reported good performance [32, 37], constructions based on minimizing the ℓ_2 loss often produce blurry outputs, as has

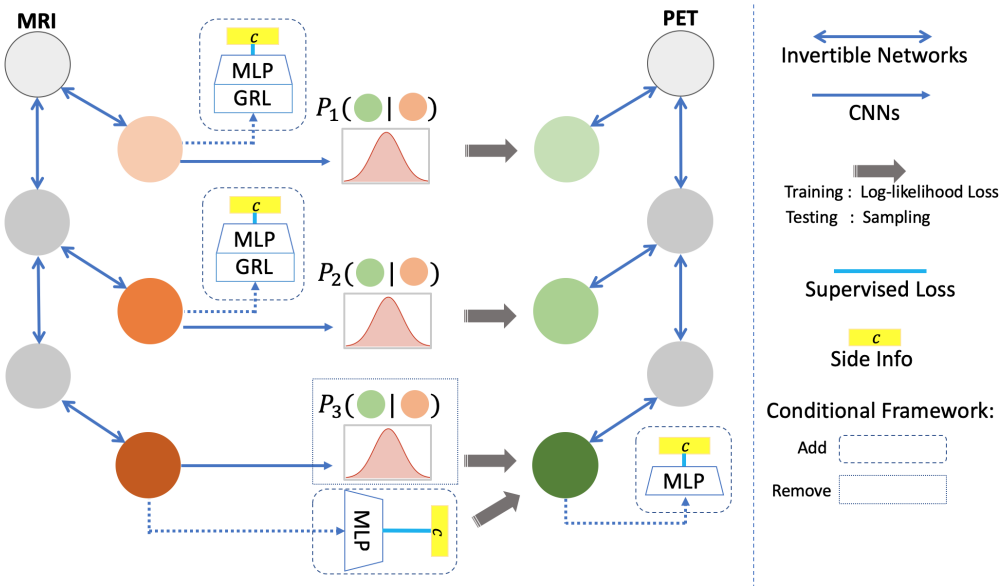

Figure 1: The DUAL-GLOW framework. For the conditional module, the dashed and dotted pieces are added and removed respectively. The colored circle represents the latent code whereas the gray one is the image or the intermediate output.

been observed in [34]. Partly due to these reasons, more recent works have investigated other generative models. Recently, one of the prominent generative models in use today, GANs [12], has seen much success in natural image synthesis [3], estimating the generative model via an adversarial process. Despite their success in generating *sharp* realistic images, GANs usually suffer from "mode collapse", that tends to produce limited sample variety [1, 4]. This issue is only compounded in medical images, where the maximal mode may simply be attributed to anatomical structure shared by most subjects. Further, sample sizes are often much smaller in medical imaging compared to computer vision, which necessitates additional adjustments to the architecture and parameters, as we found in our experiments as well.

Flow-based generative models. Another family of methods, flow-based generative models [7, 8, 21], has been proposed for variational inference and natural image generation and have only recently begun to gain attention in the computer vision community. A (*normalizing*) *flow*, proposed in [35], uses a sequence of invertible mappings to build the transformation of a probability density to approximate a posterior distribution. The flow starts with an initial variable and maps it to a variable with a simple distribution (e.g., isotropic Gaussian) by repeatedly applying the change of variable rule, similar to the inference procedure in an encoder network. For the image generation task, the initial variable is the real image with some unknown probability function. Designating a well-designed inference network, the flow will learn an accurate mapping after training. Because the flow-based model is invertible, the generation of synthetic images is straightforward by sampling from the simple distribution and "flowing" through the map in reverse. Compared with other generative models and Autoregressive Models [33], flow-based methods allow tractable and accurate log-likelihood evaluation during the training process, while also providing an efficient and exact sampling from the simple prior distribution at test time.

Where is the gap? While flow-based generative models have been successful in image synthesis, it is challenging to leverage them directly for modality transfer. It is difficult to apply existing flow-based methods to our task due to the invertibility constraint in the inference network. Apart from various technical issues, consider an intuitive example. Given an MRI, we should expect that there would be many solutions of corresponding PET images, and *vice versa*. Ideally, we prefer the model to provide a conditional distribution of the PET given an MRI – such a conditional distribution can also be meaningfully used when additional information about the subject is available.

This work. Motivated by the above considerations, we propose a novel flow-based generative model, DUAL-GLOW, for MRI-to-PET image generation. The value of our model includes explicit latent variable representations, exact and efficient latent-variable inference, and the potential for memory and computation savings through constant network size. Utilizing recent developments in flow-based generative models by [21], DUAL-GLOW is composed of two invertible inference networks and a relation CNN network, as pictured in Figure 1. We adopt the multi-scale architecture with spliting technique in [8], which can significantly reduce the computational cost and memory. The two

inference networks are built to project MRI and PET into two semantically meaningful latent spaces, respectively. The relation network is constructed to estimate the conditional distribution between paired latent codes. The foregoing properties of the DUAL-GLOW framework enable specific improvements in *modality transfer* from MRI to PET images. Sampling efficiency allows us to process and generate full 3D brain volumes.

Conditioning based on additional information. While the direct generation of PET from MRI has much practical utility, it is often also the case that a single MRI could correspond to a very different PET image – and which images are far more likely can be resolved based on additional information, such as age or disease status. However, a challenge arises due to the high correlation between the input MR image and side information: traditional conditional frameworks [21,29] cannot effectively generate meaningful images in this setting. To accurately account for this correlation, we propose a new conditional framework, see Figure 1, where two small discriminators (Multiple Layer Perceptron, MLP) are concatenated at the end of the top inference networks to faithfully extract the side information contained in the images. The remaining two discriminators concatenated at the left invertible inference network are combined with Gradient Reverse Layers (GRL), proposed in [10], to exclude the side information which exists in the latent codes except at the top-most layer. After training, sampling from the conditional distribution allows the generation of diverse and meaningful PET images. Extensive experiments show the efficiency of this exclusion architecture in the conditional framework for side information manipulation.

Contributions. This paper provides: **(1)** A novel flow-based generative model for modality transfer, DUAL-GLOW. **(2)** A complete end-to-end PET image generation from MRI for full three-dimensional volumes. **(3)** A simple extension that enables *side condition manipulation* – a practically useful property that allows assessing change as a function of age, disease status, or other covariates. **(4)** Extensive experimental analysis of the quality of PET images generated by DUAL-GLOW, indicating the potential for direct application in practice to help in the clinical evaluation of Alzheimer's disease (AD).

2. Flow-based Generative Models

We first briefly review flow-based generative models to help motivate and present our algorithm. Flow based generative models, e.g., GLOW [21], typically deal with single image generation. At a high level, these approaches set up the task as calculating the log-likelihood of an input image with an unknown distribution. Because maximizing this log-likelihood is intractable, a *flow* is set up to project the data into a new space where it is easy to compute, as summarized below.

Let $\mathbf{x}$ be an image represented as a high-dimensional random vector in the image space with an unknown true distribution $\mathbf{x} \sim p^*(\mathbf{x})$. We collect an i.i.d. dataset $\mathcal{D}$ with samples $\{\mathbf{x}^i\}_{i=1}^n$ and choose a model class $p_\theta(\mathbf{x})$ with parameters θ. Our goal is to find parameters $\hat{\theta}$ that produces $p_{\hat{\theta}}(\mathbf{x})$ to best approximate $p^*(\mathbf{x})$. This is achieved through maximization of the log-likelihood:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(\mathbf{x}^i). \tag{1}$$

In typical flow-based generative models [7, 8, 21], the *generative process* for $\mathbf{x}$ is defined in the following way:

$$\mathbf{z} \sim p_\theta(\mathbf{z}), \quad \mathbf{x} = g_\theta(\mathbf{z}), \tag{2}$$

where $\mathbf{z}$ is the latent variable and $p_\theta(\mathbf{z})$ has a (typically simple) tractable density, such as a spherical multivariate Gaussian distribution: $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$. The function $g_\theta(\cdot)$ may correspond to a rich function class, but is invertible such that given a sample $\mathbf{x}$, latent-variable inference is done by $\mathbf{z} = f_\theta(\mathbf{x}) = {g_\theta}^{-1}(\mathbf{x})$. For brevity, we will omit subscript θ from f_θ and g_θ.

We focus on functions where f is composed of a sequence of invertible transformations: $f = f_k \circ \cdots \circ f_2 \circ f_1$, where the relationship between $\mathbf{x}$ and $\mathbf{z}$ can be written as:

$$\mathbf{x} \overset{f_1}{\longleftrightarrow} \mathbf{h}_1 \overset{f_2}{\longleftrightarrow} \mathbf{h}_2 \cdots \overset{f_k}{\longleftrightarrow} \mathbf{z}. \tag{3}$$

Such a sequence of invertible transformations is also called a (normalizing) *flow* [35]. Under the change of variables rule through (2), the log probability density function of the model (1) given a sample $\mathbf{x}$ can be written as:

$$\log p_\theta(\mathbf{x}) = \log p_\theta(\mathbf{z}) + \log |det(d\mathbf{z}/d\mathbf{x})| \tag{4}$$

$$= \log p_\theta(\mathbf{z}) + \sum_{i=1}^{k} \log |det(d\mathbf{h}_i/d\mathbf{h}_{i-1})| \tag{5}$$

where we define $\mathbf{h}_0 = \mathbf{x}$ and $\mathbf{h}_k = \mathbf{z}$ for conciseness. The scalar value $\log |det(d\mathbf{h}_i/d\mathbf{h}_{i-1})|$ is the logarithm of the absolute value of the determinant of the Jacobian matrix $(d\mathbf{h}_i/d\mathbf{h}_{i-1})$, also called the log-determinant. While it may look difficult, this value can be simple to compute for certain choices of transformations, as previous explored in [7]. For the transformations $\{f_i\}_{i=1}^k$ which characterizes the flow, there are several typical settings that result in invertible functions, including actnorms, invertible 1×1 convolutions, and affine coupling layers [21]. Here we use affine coupling layers, discussed in further detail shortly. For more, details regarding these mappings we refer the reader to existing literature on flow-based models, including GLOW [21].

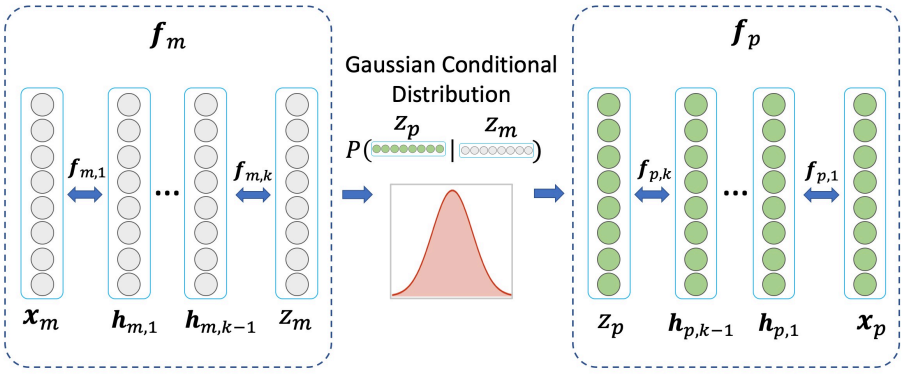

Figure 2: DUAL-GLOW for image generation.

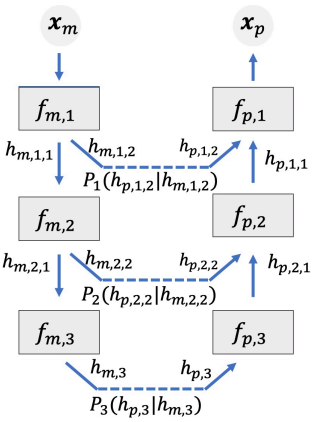

Figure 3: Spliting.

3. Deriving DUAL-GLOW

In this section, we present our DUAL-GLOW framework for inter-modality transfer. We first discuss the derivation of the conditional distribution of a PET image given an MR image and then provide strategies for efficient calculation of its log-likelihood. Then, we introduce the construction of the invertible flow and show the calculation for the Jacobian matrix. Next, we build the hierarchical architecture for our DUAL-GLOW framework, which greatly reduces the computational cost compared to a flat structure. Finally, the conditional structure for side information manipulation is derived with additional discriminators.

Log-Likelihood of the conditional Distribution. Let the data corresponding to the MR and PET images be denoted as $\mathcal{D}_m$ and $\mathcal{D}_p$. From a dataset $\mathcal{D}_m = \{\mathbf{x}_m^i\}_{i=1}^n$, we are interested in generating images which have the same properties as images in the dataset $\mathcal{D}_p = \{\mathbf{x}_p^i\}_{i=1}^n$. In our DUAL-GLOW model, we assume that there exists a flow-based invertible function f_p which maps the PET image $\mathbf{x}_p$ to $\mathbf{z}_p = f_p(\mathbf{x}_p)$ and a flow-based invertible function f_m which maps the MR image $\mathbf{x}_m$ to $\mathbf{z}_m = f_m(\mathbf{x}_m)$. The latent variables $\mathbf{z}_p$ and $\mathbf{z}_m$ help set up a conditional probability $p_\theta(\mathbf{z}_p|\mathbf{z}_m)$, given by

$$p_\theta(\mathbf{z}_p|\mathbf{z}_m) = \mathcal{N}(\mathbf{z}_p; \mu_\theta(\mathbf{z}_m), \sigma_\theta(\mathbf{z}_m)) \tag{6}$$

The full mapping composed of f_p, f_m, μ_θ and σ_θ formulates our DUAL-GLOW framework:

$$\mathbf{x}_m \overset{f_m}{\longleftrightarrow} \mathbf{z}_m \overset{\mu_\theta,\sigma_\theta}{\longrightarrow} \mathbf{z}_p \overset{{f_p}^{-1}}{\longleftrightarrow} \mathbf{x}_p, \tag{7}$$

see Figure 2. The invertible functions f_p and f_m are designed as flow-based invertible functions. The mean function μ_θ and the covariance function σ_θ for $p_\theta(\mathbf{z}_p|\mathbf{z}_m)$ are assumed to be specified by neural networks. In this generating process, our goal is to maximize the log conditional probability $p_\theta(\mathbf{x}_p|\mathbf{x}_m)$. By the change of variable rule, we have that

$$\begin{aligned}
\log p_\theta(\mathbf{x}_p|\mathbf{x}_m) &= \log\left(p_\theta(\mathbf{x}_p, \mathbf{x}_m)/p_\theta(\mathbf{x}_m)\right) && (8)\\
&= \log p_\theta(\mathbf{z}_p|\mathbf{z}_m) + \log\left(\frac{|det(d(\mathbf{z}_p, \mathbf{z}_m)/d(\mathbf{x}_p, \mathbf{x}_m))|}{|det(d\mathbf{z}_m/d\mathbf{x}_m)|}\right) && (9)\\
&= \log p_\theta(\mathbf{z}_p|\mathbf{z}_m) + \log\left(|det(d\mathbf{z}_p/d\mathbf{x}_p)|\right). && (10)
\end{aligned}$$

Note that the Jacobian $d(\mathbf{z}_p, \mathbf{z}_m)/d(\mathbf{x}_p, \mathbf{x}_m)$ in (9) is, in fact, a block matrix

$$\frac{d(\mathbf{z}_p, \mathbf{z}_m)}{d(\mathbf{x}_p, \mathbf{x}_m)} = \begin{bmatrix} d\mathbf{z}_p/d\mathbf{x}_p & 0 \\ 0 & d\mathbf{z}_m/d\mathbf{x}_m \end{bmatrix}. \tag{11}$$

Recall that calculating the deteriminant of such a matrix is straightforward (see [38]), which leads directly to (10).

Without any regularization, maximizing such a conditional probability can make the optimization hard. Therefore, we may add a regularizer by controlling the marginal distribution $p_\theta(\mathbf{z}_m)$, which leads to our objective function

$$\begin{aligned}
\max_{f_m, f_p, \mu_\theta, \sigma_\theta} & \log p_\theta(\mathbf{z}_p|\mathbf{z}_m) + \log(|det(\frac{d\mathbf{z}_p}{d\mathbf{x}_p})|) + \lambda \log p_\theta(\mathbf{x}_m) \\
&= \log p_\theta(f_p(\mathbf{x}_p)|f_m(\mathbf{x}_m)) + \log(|det(\frac{d\mathbf{z}_p}{d\mathbf{x}_p})|) \\
&\quad + \lambda \log(p_\theta(f_m(\mathbf{x}_m))|det(\frac{d\mathbf{z}_m}{d\mathbf{x}_m})|), \qquad (12)
\end{aligned}$$

where λ is a hyperparameter, $p_\theta(\mathbf{z}_m) = \mathcal{N}(\mathbf{z}_m; 0, \mathbf{I})$ and $p_\theta(\mathbf{z}_p|\mathbf{z}_m) = \mathcal{N}(\mathbf{z}_p; \mu_\theta(\mathbf{z}_m), \sigma_\theta(\mathbf{z}_m))$.

Interestingly, compared to GLOW, our model does **not** introduce much additional complexity in computation. Let us see why. First, the marginal distribution $p_\theta(\mathbf{z})$ in GLOW is replaced by $p_\theta(\mathbf{z}_p|\mathbf{z}_m)$ and $p_\theta(\mathbf{z}_m)$, which still has a simple and tractable density. Second, instead of one flow-based invertible function in GLOW, our DUAL-GLOW has two

flow-based invertible functions f_p, f_m. Those functions are setup in parallel based on (12), extending the model size by a constant factor.

Flow-based Invertible Functions. In our work, we use an affine coupling layer to design the flows for the invertible functions f_p and f_m. Before proceeding to the details, we omit subscripts p and m to simplify notations in this subsection. The invertible function f is composed of a sequence of transformations $f = f_k \circ \cdots \circ f_2 \circ f_1$, as introduced in (3). In DUAL-GLOW, $\{f_i\}_{i=1}^k$ are designed by using the affine coupling layer [8] following these equations:

$$\mathbf{h}_i = f_i(\mathbf{h}_{i-1}) \Leftrightarrow \begin{cases} \mathbf{h}_{i;1:d_1} = \mathbf{h}_{i-1;1:d_1} \\ \mathbf{h}_{i;d_1+1:d_1+d_2} = (\mathbf{h}_{i-1;d_1+1:d_1+d_2} \odot \exp(\mathrm{s}(\mathbf{h}_{i-1;1:d_1})) \\ \qquad\qquad + \ \mathrm{t}(\mathbf{h}_{i-1;1:d_1}), \end{cases} \tag{13}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{h}_i \in R^{d_1+d_2}$, $\mathbf{h}_{i;1:d_1}$ the first d_1 dimensions of $\mathbf{h}_i$, and $\mathbf{h}_{i;d_1+1:d_1+d_2}$ the remaining d_2 dimensions of $\mathbf{h}_i$. The functions $s(\cdot)$ and $t(\cdot)$ are nonlinear transformations where it makes sense to use deep convolutional neural networks (DCNNs). This construction makes the function f invertible. To see this, we can easily write the inverse function ${f_i}^{-1}$ for f_i as

$$\mathbf{h}_{i-1} = (f_i)^{-1}(\mathbf{h}_i) \Leftrightarrow \begin{cases} \mathbf{h}_{i-1;1:d_1} = \mathbf{h}_{i;1:d_1} \\ \mathbf{h}_{i-1;d_1+1:d_1+d_2} = (\mathbf{h}_{i;d_1+1:d_1+d_2} - \mathrm{t}(\mathbf{h}_{i;1:d_1})) \\ \qquad\qquad \odot \ \exp(-\mathrm{s}(\mathbf{h}_{i;1:d_1})). \end{cases} \tag{14}$$

In addition to invertibility, this structure also tells us that the $\log(|det(d\mathbf{z}_p/d\mathbf{x}_p)|)$ term in our objective (12) has a simple and tractable form. Computing the Jacobian, we have:

$$\frac{\partial f_i(\mathbf{h}_{i-1})}{\partial \mathbf{h}_{i-1}} = \begin{bmatrix} \mathrm{I}_{1:d_1} & 0 \\ \frac{\partial f_{i;d_1+1:d_1+d_2}}{\partial \mathbf{h}_{i-1;1:d_1}} & \mathrm{diag}\big(\exp\big(\mathrm{s}(\mathbf{h}_{i-1;1:d_1})\big)\big) \end{bmatrix}, \tag{15}$$

where $I_{1:d_1} \in R^{d_1 \times d_1}$ is an identity matrix. Therefore,

$$\begin{aligned} \log(|det(d\mathbf{z}_p/d\mathbf{x}_p)|) &= \sum_{i=1}^{k} \log(|det(d\mathbf{h}_i/d\mathbf{h}_{i-1})|) \\ &= \sum_{i=1}^{k} \log(|det(\mathrm{diag}\big(\exp\big(\mathrm{s}(\mathbf{h}_{i-1;1:d_1})\big)\big))|) \\ &= \sum_{i=1}^{k} \log \left| \exp\left(\sum_{j=1}^{d_1} \mathrm{s}(\mathbf{h}_{i-1;j}) \right) \right| \end{aligned}$$

which can be computed easily and efficiently, requiring no on-the-fly matrix inversions [21].

Efficiency from Hierarchical Structure. The flow $f = f_k \circ \cdots f_2 \circ f_1$ can be viewed as a hierarchical structure. For the two datasets $\mathcal{D}_m = \{\mathbf{x}_m^i\}_{i=1}^n$ and $\mathcal{D}_p = \{\mathbf{x}_p^i\}_{i=1}^n$, it is computationally expensive to make all features of all samples go through the entire flow. Following implementation strategies in previous flow-based models, we use the splitting technique to speed up DUAL-GLOW in practice, see Figure 3. When a sample $\mathbf{x}$ reaches the i-th transformation f_i in the flow as $\mathbf{h}_{i-1}$, we split $\mathbf{h}_{i-1}$ in two parts $\mathbf{h}_{i-1,1}$ and $\mathbf{h}_{i-1,2}$, and take only one part $\mathbf{h}_{i-1,1}$ through f_i to become $\mathbf{h}_i = f_i(\mathbf{h}_{i-1,1})$. The other part $\mathbf{h}_{i-1,2}$ is taken out from the flow without further transformation. Finally, all those split parts $\{\mathbf{h}_{i,2}\}_{i=1}^{k-1}$ and the top-most $\mathbf{h}_k$ are concatenated together to form $\mathbf{z}$. By using this splitting technique in the flow hierarchy, the part leaving the flow "early" goes through fewer transformations. As discussed in GLOW and previous flow-based models, each transformation f_i is usually rich enough that splitting saves computation without losing much quality in practice. We provide the computational complexity in the appendix. Additionally, this hierarchical representation enables a more succinct extension to allow side information manipulation.

How to condition based on side information? As stated above, additional covariates should influence the PET image we generate, even with a very similar MRI. A key assumption in many conditional side information frameworks is that these two inputs (the input MR and the covariate) are independent of each other. Clearly, however, there exists a high correlation between MRI and side information such as age or gender or disease status. In order to effectively incorporate this into our model, it is necessary to disentangle the side information from the intrinsic properties encoded in the latent representation $\mathbf{z}_m$ of the MR image.

Let c denote the side information, typically a high-level semantic label (age, sex, disease status, genotype). In this case, we expect that the effect of this side information would be at a high level in relation to individual image voxels. As such, we expect that only the highest level of DUAL-GLOW should be affected by this. The latent variables $\mathbf{z}_p$ should be conditioned on side variable c and $\mathbf{z}_m = \{\mathbf{h}_{i,2}\}_{i=1}^k$ except $\mathbf{h}_k$. Thus, we can rewrite the conditional probability in (12) by adding c:

$$\begin{aligned} \max_{f_m, f_p, \mu_\theta, \sigma_\theta} & \log p_\theta(\mathbf{z}_p | \mathbf{z}'_m, c) \\ & + \log(|det(\frac{d\mathbf{z}_p}{d\mathbf{x}_p})|) + \lambda \log p_\theta(\mathbf{x}_m), \end{aligned} \tag{16}$$

where $\mathbf{z}'_m = \{\mathbf{h}_{i,2}\}_{i=1}^{k-1}$ is independent on c, and

$$p_\theta(\mathbf{z}_p | \mathbf{z}'_m, c) = \mathcal{N}(\mathbf{z}_p; \mu_\theta(\mathbf{z}'_m, c), \sigma_\theta(\mathbf{z}'_m, c)). \tag{17}$$

To disentangle the latent representation $\mathbf{z}'_m$ and exclude the side information in $\mathbf{z}'_m$, we leverage the well-designed conditional framework composed of both flow and discriminators. Specifically, the condition framework tries to exclude

the side information from $\{\mathbf{h}_{i,2}\}_{i=1}^{k-1}$ and keep it in $\mathbf{h}_k$ at the top level during training time. To achieve this, we concatenate a simple discriminator for each $\{f_i\}_{i=1}^{k-1}$ and add a Gradient Reversal Layer (GRL), introduced in [10], at the beginning of the network. These classifiers are used for distinguishing the side information in a supervised way. The GRL acts as the identity function during the forward-propagation and reverses the gradient in back-propagation. Therefore, minimizing the classification loss in these classifiers is equivalent to pushing the model to exclude the information gained by this side information, leading to the exclusive representation $\mathbf{z}'_m = \{\mathbf{h}_{i,2}\}_{i=1}^{k-1}$. We also add a classifier *without* GRL at the top level of f_m, f_p that explicitly preserves this side information at the highest level.

Finally, the objective is the log-likelihood loss in (16) plus the classification losses, which can be jointly optimized by the popular optimizer AdaMax [20]. The gradient is calculated in a memory efficient way inspired by [5]. After training the conditional framework, we achieve PET image generation influenced **both** by MRI and side information.

4. Experiments

We evaluate the model's efficacy on the ADNI dataset both against ground truth images and for downstream applications. We conduct extensive quantitative experiments which show that DUAL-GLOW outperforms the baseline method consistently. Our generated PET images show desirable clinically meaningful properties which is relevant for their potential use in Alzheimer's Disease diagnosis. The conditional framework also shows promise in tracking hypometabolism as a function of age.

4.1. ADNI Dataset

Data. The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides a large database of studies directly aimed at understanding the development and pathology of Alzheimer's Disease. Subjects are diagnosed as 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). FDG-PET and T1-weighted MRIs were obtained from ADNI, and pairs were constructed by matching images with the same subject ID and similar acquisition dates.

Preprocessing. Images were processed using SPM12 [2].

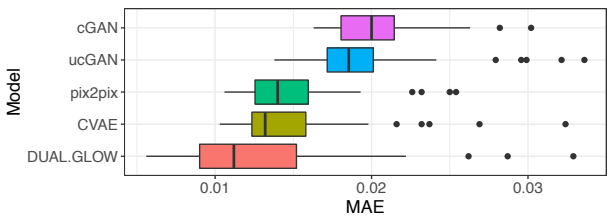

Figure 4: Box plot of MAE metrics for different methods.

First, PET images were aligned to the paired MRI using coregistration. Next, MR images were nonlinearly mapped to the MNI152 template. Finally, PET images were mapped to the standard MNI space using the same forward warping identified in the MR segmentation step. Voxel size was fixed for all volumes to $1.5 \times 1.5 \times 1.5 mm^3$, and the final volume size obtained for both MR and PET images was $64 \times 96 \times 64$. Through this workflow, we finally obtain 806 MRI/PET clean pairs. The demographics of the dataset are provided in the appendix. In the following experiments, we randomly select 726 subjects as the training data and the remaining 80 as testing within a 10-fold evaluation scheme.

Framework Details. The DUAL-GLOW architecture outlined above was trained using Nvidia V100 GPUs with Tensorflow. There are 4 "levels" in our invertible network, each containing 16 affine coupling layers. The nonlinear operators $\mathrm{s}(\cdot)$ and $\mathrm{t}(\cdot)$ are small networks with three 3D convolutional layers. For the hierarchical correction learning network, we split the hidden codes of the output of the first three modules in the invertible network and design four 3D convolutional networks for all latent codes. For the conditional framework case, we concatenate the five discriminators to the tail of all four levels of the MRI inference network and the top-most level of the PET inference network. The GRL is added between the inference network and the first three discriminators. The hyperparameter λ is the regularizer and set to 0.001. For all classification losses, we set the weight to 0.01. The model was trained using the AdamMax optimizer with an initial learning rate set to 0.001 and exponential decay rates 0.9 for the moment estimates. We train the model for 90 epochs. Our implementation is available at `https://github.com/haolsun/dual-glow`.

4.2. Generated versus Ground Truth consistency

We begin our model evaluation by comparing outputs from our model to 4 state-of-the-art methods used previously for similar image generation tasks, conditional GANs (cGANs) [32], cGANs with U-Net architecture (UcGAN) [36], Conditional VAE (C-VAE) [9, 39], pix2pix [16]. Additional experimental setup details are in the appendix. We compare using commonly-used quantitative measures computed over the held out testing data. These include Mean Absolute Error (MAE), Correlation Coefficients (CorCoef), Peak Signal-to-Noise Ratio (PSNR), and Structure Similarity Index (SSIM). For Cor_Coef, PSNR and SSIM, higher values indicate better generation of PET images. For MAE, the lower the value, the better is the generation. As seen in Table 1 and Figure 4, our model competes favorably against other methods.

Figure 5 shows test images generated after 90 epochs for Cognitively Normal and Alzheimer's Disease individuals. Qualitatively, not only is the model able to accurately re-

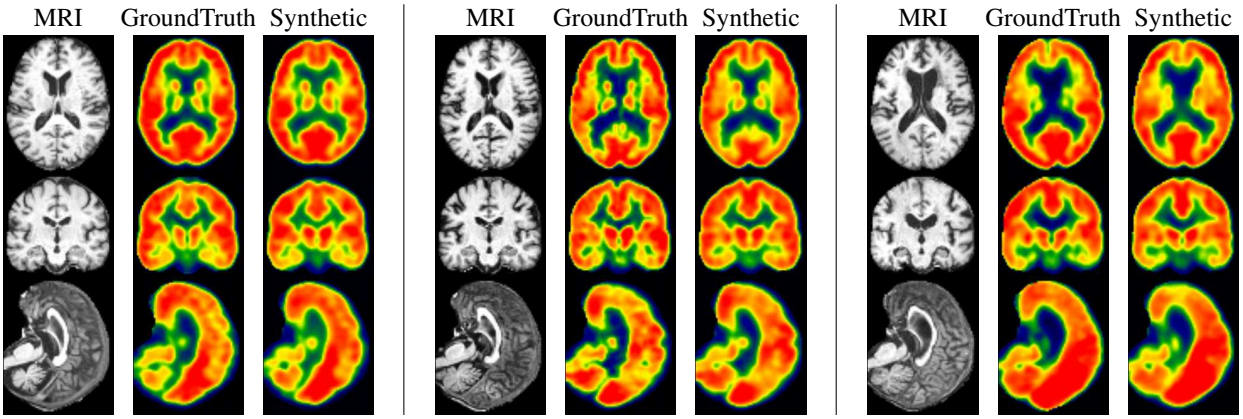

Figure 5: **Synthetic images are meaningful for subjects in both extremes of disease spectrum**. Left: CN. Middle: MCI. Right: AD. The generated PET images show consistency of hypometabolism (less red, more yellow) with the ground truth image. (Best viewed in color; montages shown in the appendix).

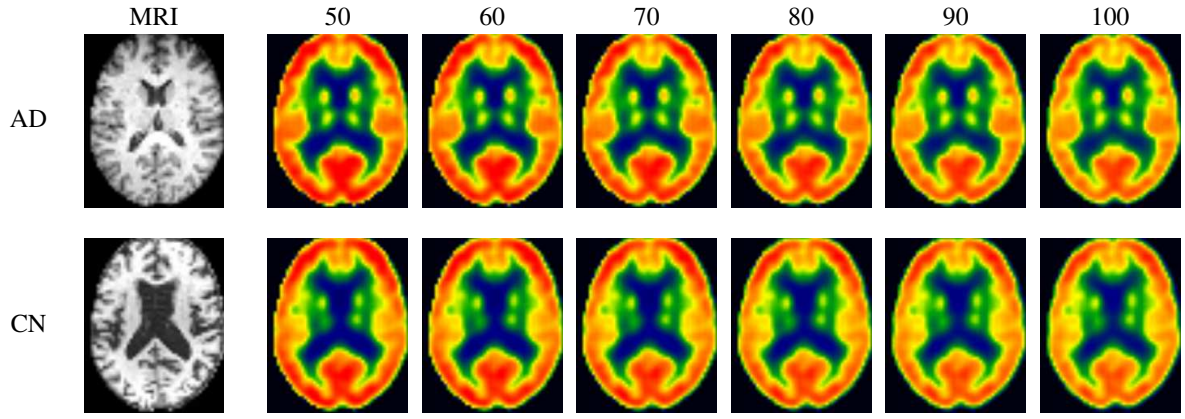

Figure 6: **Conditioning on age should yield generated images that show increased hypometabolism with age.** These are representative results from our PET generation as a function of age. As we scan left to right, we indeed see a decrease in metabolism (less red, more yellow) which is completely consistent with what we would expect in aging. (Best viewed in color; montages shown in the appendix).

construct large scale anatomical structures but it is also able to identify minute, sharp boundaries between gray matter and white matter. While here we focus on data from individuals with a clear progression of Alzheimer's disease from those who are clearly cognitively healthy, in preclinical cohorts where disease signal may be weak, accurately constructing finer-grained details may be critical in identifying those who may be undergoing neurodegeneration due to dementia. More results are shown in the appendix.

4.3. Scientific Evaluation of Generation

As we saw above, our method is able to learn the modality mapping from MRI to PET. However, often image acquisition is used as a means to an end: typically towards disease diagnosis or informed preventative care. While the generated images may seem computationally and visually coherent, it is important that the images generated add some value towards these downstream analyses.

We also evaluate the **generated** PET images for disease prediction and classification. Using the AAL atlas, we obtain all 116 ROIs via atlas-based segmentation [45] and use the mean intensity of each as image features. A support vector machine (SVM) is trained with the standard RBF kernel (e.g., see [14]) to predict binary disease status (Normal, EMCI, SMC vs. MCI, LMCI, AD) for both the ground truth and the generated images. The SVM trained on generated images achieves comparable accuracy and false positive/negative rates (Table 2), suggesting that the generated images contain sufficient discriminative signal for disease diagnosis.

Adjusting for Age with Conditioning. The conditional framework naturally allows us to evaluate potential developing pathology as an individual ages. Training the full conditional DUAL-GLOW model, we use ground truth "side" information (age) as the conditioning variable de-

Table 1: Quantitative comparison results on 10-fold cross-validation.

METHOD	cGAN	UcGAN	C-VAE	pix2pix	DUAL-GLOW
CorCoef	0.956	0.963	**0.980**	0.967	0.975
PSNR	27.37 ± 2.07	27.84 ± 1.23	28.69 ± 2.06	27.54 ± 1.95	$\mathbf{29.56 \pm 2.66}$
SSIM	0.761 ± 0.08	0.780 ± 0.06	0.817 ± 0.06	0.783 ± 0.05	$\mathbf{0.898 \pm 0.06}$

Table 2: Validation on the ground truth and synthetic images for the AD/CN classification.

	Ground Truth	Synthetic
Accuracy	94%	91%
False Negative Rate	6%	6%
False Positive Rate	0%	3%

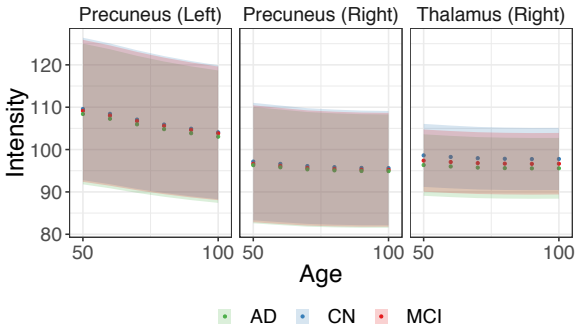

Figure 7: The mean intensity with 95% standard deviation bands of 3 ROIs with the change of age for all test subjects. The clear downward trend reflects expected hypometabolism as a function of age.

scribed above. Figure 6 shows the continuous change in the 3 generated images given various age labels for the same MRI. The original image (left) is at age 50, and as we increase age from 60 to 100, increased hypometabolism becomes clear. To quantitatively evaluate our conditional framework, we plot the mean intensity value of a few key ROIs. As we see in Figure 7, the mean intensity values show a downward trend with age, as expected. While there is a clear 'shift' between AD, MCI, and CN subjects (blue dots lie above red dots, etc.), the wide variance bands indicate a larger sample size may be necessary to derive statistically sound conclusions (e.g., regarding group differences). Additional results and details can be found in the appendix.

4.4. Other Potential Applications

While not a key focus of our work, to show the model's generality on visually familiar images we directly test DUAL-GLOW's ability to generate images on a standard computer vision modality transfer task. Using the UT-Zap50K dataset [47, 48] of shoe images, we construct HED [46] edge images as "sketches", similar to [16]. We aim to learn a mapping from sketch to shoe. We also create a cartoon face dataset based on CelebA [26] and train our model to generate a realistic image from the cartoon face. Fig. 8

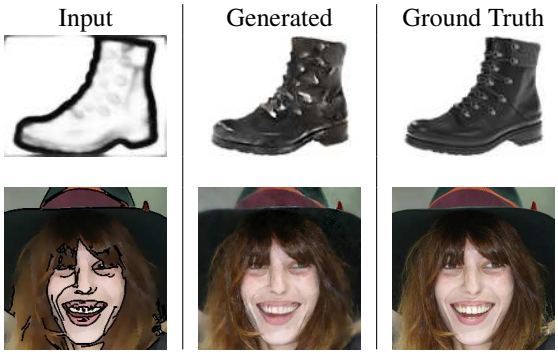

Figure 8: Sample generation using DUAL-GLOW. The first row: UT-Zap50K dataset. The second row: CartoonFace dataset.

shows the results of applying our model (and ground truth). Clearly, more specialized networks designed for such a task will yield more striking results, but these experiments suggest that the framework is general and applicable in additional settings. These results are available on the project homepage and in the appendix.

5. Conclusions

We propose a flow-based generative model, DUAL-GLOW, for inter-modality transformation in medical imaging. The model allows for end-to-end PET image generation from MRI for full three-dimensional volumes, and takes advantage of explicitly characterizing the conditional distribution of one modality given the other. While inter-modality transfer has been reported using GANs, we present improved results along with the ability to condition the output easily. Applied to the ADNI dataset, we are able to generate sharp synthetic PET images that are scientifically meaningful. Standard correlation and classification analysis demonstrates the potential of generated PET in diagnosing Alzheimer's Disease, and the conditional side information framework is promising for assessing the change of spatial metabolism with age.

Acknowledgments. HS was supported by Natural Science Foundation of China (Grant No. 61876098, 61573219) and scholarships from China Scholarship Council (CSC). Research was also supported in part by R01 AG040396, R01 EB022883, NSF CAREER award RI 1252725, UW Draper Technology Innovation Fund (TIF) award, UW CPCP AI117924 and a NIH predoctoral fellowship to RM via T32 LM012413.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] John Ashburner, Gareth Barnes, C Chen, Jean Daunizeau, Guillaume Flandin, Karl Friston, Stefan Kiebel, James Kilner, Vladimir Litvak, Rosalyn Moran, et al. Spm12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, 2014.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

[5] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[7] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[9] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[11] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, pages 1294–1305, 2018.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] Xiao Han. Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical physics*, 44(4):1408–1419, 2017.

[14] Chris Hinrichs, Vikas Singh, Guofan Xu, and Sterling Johnson. Mkl for robust multi-modality ad classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 786–794. Springer, 2009.

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, pages 172–189, 2018.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[17] Jiayin Kang, Yaozong Gao, Feng Shi, David S Lalush, Weili Lin, and Dinggang Shen. Prediction of standard-dose brain pet image by using mri and low-dose brain [18f] fdg pet images. *Medical physics*, 42(9):5301–5309, 2015.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. JMLR. org, 2017.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10235–10244, 2018.

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, pages 35–51, 2018.

[23] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer, 2014.

[24] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Computer Vision and Pattern Recognition*, pages 5524–5532, 2018.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, pages 3730–3738, 2015.

[27] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *International Conference on Learning Representations*, 2019.

[28] Matteo Maspero, Mark HF Savenije, Anna M Dinkla, Peter R Seevinck, Martijn PW Intven, Ina M Jurgenliemk-Schulz, Linda GW Kerkmeijer, and Cornelis AT van den Berg. Dose evaluation of fast synthetic-ct generation using a generative adversarial network for general pelvis mr-only radiotherapy. *Physics in Medicine & Biology*, 63(18):185001, 2018.

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

[30] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.

[31] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[32] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

[33] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[34] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Tao Zhou, Yong Xia, and Dinggang Shen. Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimers disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 455–463. Springer, 2018.

[35] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[36] 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.

[37] Apoorva Sikka, Skand Vishwanath Peri, and Deepti R Bathula. Mri to fdg-pet: Cross-modal synthesis using 3d u-net for multi-modal alzheimers classification. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 80–89. Springer, 2018.

[38] John R Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, 2000.

[39] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[40] Haoliang Sun, Xiantong Zhen, Chris Bailey, Parham Rasoulinejad, Yilong Yin, and Shuo Li. Direct estimation of spinal cobb angles by structured multi-output regression. In *International Conference on Information Processing in Medical Imaging*, pages 529–540. Springer, 2017.

[41] Haoliang Sun, Xiantong Zhen, Yuanjie Zheng, Gongping Yang, Yilong Yin, and Shuo Li. Learning deep match kernels for image-set classification. In *Computer Vision and Pattern Recognition*, pages 3307–3316, 2017.

[42] Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27(8):4066–4079, 2018.

[43] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2018.

[44] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.

[45] Minjie Wu, Caterina Rosano, Pilar Lopez-Garcia, Cameron S Carter, and Howard J Aizenstein. Optimum template selection for atlas-based segmentation. *NeuroImage*, 34(4):1612–1618, 2007.

[46] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *International Conference on Computer Vision*, pages 1395–1403, 2015.

[47] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition*, Jun 2014.

[48] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision*, Oct 2017.

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017.

[50] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.