Deep Adversarial Learning for Multi-Modality Missing Data Completion

Lei Cai
Washington State University
Pullman, WA, USA
lei.cai@wsu.edu

Zhengyang Wang
Washington State University
Pullman, WA, USA
zwang6@eecs.wsu.edu

Hongyang Gao
Washington State University
Pullman, WA, USA
hongyang.gao@wsu.edu

Dinggang Shen
University of North Carolina
Chapel Hill, NC, USA
dgshen@med.unc.edu

Shuiwang Ji
Washington State University
Pullman, WA, USA
sji@eecs.wsu.edu

ABSTRACT

Multi-modality data are widely used in clinical applications, such as tumor detection and brain disease diagnosis. Different modalities can usually provide complementary information, which commonly leads to improved performance. However, some modalities are commonly missing for some subjects due to various technical and practical reasons. As a result, multi-modality data are usually incomplete, raising the multi-modality missing data completion problem. In this work, we formulate the problem as a conditional image generation task and propose an encoder-decoder deep neural network to tackle this problem. Specifically, the model takes the existing modality as input and generates the missing modality. By employing an auxiliary adversarial loss, our model is able to generate high-quality missing modality images. At the same time, we propose to incorporate the available category information of subjects in training to enable the model to generate more informative images. We evaluate our method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, where positron emission tomography (PET) modalities are missing. Experimental results show that the trained network can generate high-quality PET modalities based on existing magnetic resonance imaging (MRI) modalities, and provide complementary information to improve the detection and tracking of the Alzheimer's disease. Our results also show that the proposed methods generate higher quality images than baseline methods as measured by various image quality statistics.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Life and medical sciences**; **Imaging**;

KEYWORDS

Deep learning, adversarial loss function, missing data completion, disease diagnosis

ACM Reference Format:
Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep Adversarial Learning for Multi-Modality Missing Data Completion. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219963

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219963

1 INTRODUCTION

Many clinical applications[35, 37], such as tumor detection and brain disease diagnosis [1, 21, 30, 32], require high-quality multi-modality data in order to achieve good diagnostic results, since different modalities of a subject provide complementary information. While standardized methods for clinical tests have been developed to collect multi-modality data, there are some practical concerns in the process of obtaining some important and informative modalities. For example, the positron emission tomography (PET) modality is often used to reveal metabolic information, in addition to anatomical details provided by other common modalities like magnetic resonance imaging (MRI) modality. However, to obtain PET images of diagnostic quality, a living subject needs to take an injection of a radioactive tracer. It raises the risk of radioactive exposure, resulting in potential harm to one's health. Therefore, the PET scan is rejected in some cases, where the data of a subject are incomplete with missing PET modality. While completely safe methods have not been developed, it is desired to perform the multi-modality missing data completion, where one can generate missing modalities based on available modalities. In this work, we explore a deep learning [18] solution to this problem.

Since multi-modality data are in a 3D imaging format, we formulate the missing data completion [33, 34] problem as a conditional image generation task; that is, we aim at generating missing modality images conditioned on existing modality images. As multi-modality data are collected from the same subject, there must be some underlying relationships between modalities, although they focus on different information. Therefore, the task is feasible if one can capture the relationships and estimate a mapping from existing modalities to missing ones. Deep learning has achieved

great success in such conditional image generation tasks, like image super-resolution [20, 25], image-to-image translation [7, 14, 38], and video prediction [23]. In addition, because both the inputs and outputs of such tasks are images, usually of the same spatial size, other image tasks of structured outputs such as image segmentation [2, 4, 6, 8, 9, 22, 27] are also related.

With the fast development of deep models on these tasks, we propose an efficient and effective model for multi-modality missing data completion. However, it is worth noting that there is a major difference in our problem. For all of the relevant tasks above, the models are trained on two types of images, used as inputs and outputs, respectively. During inference, one type of images is predicted given the other. In the multi-modality missing data completion task, another kind of information, the category labels of subjects, is available during training. Such labels cannot be used as inputs to a model, as they are not accessible when performing prediction. Nor should they be outputs, because they are not what we aim to predict. Yet they provide useful information to guide the missing data completion process. How to take an advantage of these labels during training without affecting the inference phase remains challenging. We propose an elegant approach to utilize them in our model, based on our understanding of generative adversarial networks [10].

We evaluate our model on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, where the MRI modality data are given and the PET modality data are missing. Experimental results show that the proposed network can generate high-quality PET modality data based on the MRI modality data, and provides complementary information to improve the detection and tracking of the Alzheimer's disease. Our results also show that the proposed methods generate higher quality images than baseline methods as measured by various image quality statistics.

2 RELATED WORK

The multi-modality missing data completion problem was previously addressed in [21]. The authors employed a 3D convolutional neural network (CNN) architecture with only two hidden layers. There are two main drawbacks in this approach. First, the model only has two hidden layers without any downsampling layers, resulting in that any voxel in the final feature maps only incorporates information from a receptive field of a limited size in the inputs. In practice, such behavior commonly leads to poor performance when the spatial size of inputs is large. To overcome this, cropped patches of a small size ($15 \times 15 \times 15 \times 1$) were applied as inputs to their model, which leads to the second problem. Since the CNN architecture uses convolutional operations without padding and performs no upsampling, the outputs become even smaller, with a size of $3 \times 3 \times 3 \times 1$. However, the required outputs usually have a much larger spatial size. Therefore, during inference, they had to scan the model on the full inputs and then concatenated all the outputs together to form a final prediction. For instance, the spatial size of the missing PET modalities in the ADNI database is $64 \times 64 \times 64 \times 1$. Nearly ten thousand times of scanning is needed at least, which results in excessively slow inference. In addition, inconsistency happens near the edges of different output patches, hindering the performance of their model.

Similar problems were observed in image tasks of structured outputs such as image segmentation. To address this, an encoder-decoder architecture, with different upsampling layers, was developed [2, 4, 6, 8, 9, 22, 27]. Basically, by adding upsampling layers, downsampling layers that expand the receptive fields are enabled in the encoder, allowing inputs of a larger size. Moreover, the outputs can have the same size as the inputs, making the inference efficient.

In terms of image generation tasks, generative adversarial networks (GANs) were proposed by [10] and achieved impressive results. The GAN framework consists of a generator network and a discriminator network. The generator maps from latent representations to images while the discriminator is used to distinguish generated images from images from the dataset. By incorporating conditional information in the latent representations, GANs can be easily generalized to conditional GANs [3, 24, 36]. Most recent studies on conditional image generation tasks [7, 14, 20, 38] employed the encoder-decoder architecture as the generator network, which encodes the conditional information to latent representations.

To understand the GAN framework, we prefer thinking of it as a generator network with an adversarial loss function instead of two networks. Note that the discriminator network is only involved during training. When performing inference, the generator network alone is applied. Thus, the discriminator can be considered as a loss function, which is trainable and differs from a regular fixed one in regular deep learning models. With such an interpretation of GANs, we are free to use extra available information in the discriminator during training to provide a better loss function, without any influence on the inference phase.

In recent studies, training the generator network with extra loss functions in addition to the adversarial loss was found beneficial [5, 7, 14, 20, 23, 25, 38]. It was pointed out that the adversarial loss function encourages generating sharp images while the generated images can be significantly different from true ones. In contrast, regular content loss functions, such as $\mathcal{L}_1$ and $\mathcal{L}_2$ loss, are able to force generating images of similar appearances to those in the dataset, but suffer from the blurring problem. As a result, by finding an appropriate tradeoff among various loss functions, it is possible to train the same model but achieve improved performance.

In this work, we propose a 3D encoder-decoder model with multiple loss functions of different functionalities for the multi-modality missing data completion problem. Our model incorporates all available information during training and generates high-quality modality data in an efficient way.

3 METHODS

3.1 Problem Formulation

In multi-modality missing data completion problems, we are given a dataset of subjects, in which a subject I is composed of two modalities $\{x, y\}$ and a corresponding category label ℓ. We assume the modality x is available for all subjects, while modality y is available for only a portion of the subjects. The training set consists of subjects with both x and y as $\{x_i, y_i, \ell_i\}_{i=1}^{N}$ while the test set consists of subjects with only x. The multi-modality missing data completion task aims to predict y for subjects in the test set. In order to achieve this, we attempt to capture the mapping from x to y by developing a model on the training set. With the development of

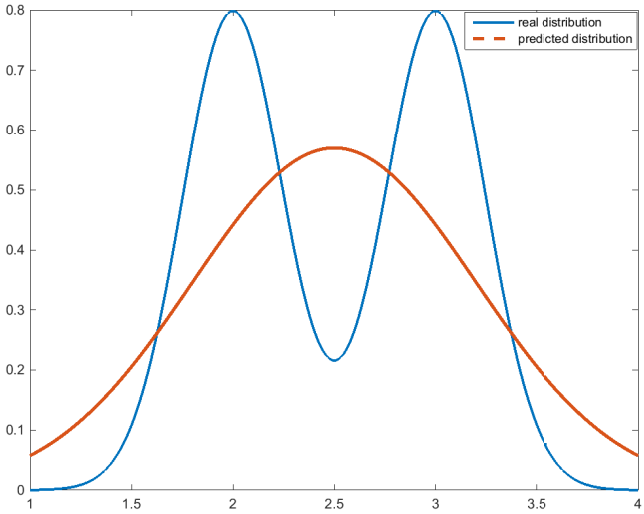

Figure 1: Illustration of the blurry effect of $\mathcal{L}_{MSE}$ on data following a mixture of two Gaussians. The blue line represents the mixture of two Gaussians. The red line represents the distribution estimated by $\mathcal{L}_{MSE}$

deep learning[17, 19], we propose to model this relationship using a deep neural network. Specifically, our goal is to train a generator network $G(x;\theta)$ to estimate the missing modality y where G is parameterized by θ. From a probabilistic perspective, suppose y is drawn from an underlying distribution $p_Y(y)$ and x is drawn from $p_X(x)$. The model G estimates a conditional probability $p_G(y|x;\theta)$ to approximate $p_Y(y)$. Note that in this setting, the category label ℓ is not used for prediction. We will discuss an approach to incorporate the category information in prediction in Section 3.4.

3.2 Content Loss

In order to train the generator network G, a content loss function is employed to encourage $G(x;\theta)$ to be close to y. The most straightforward way to achieve this is to minimize the Euclidean distance between them, resulting in the mean squared error (MSE) loss $\mathcal{L}_{MSE}$ defined as

$$\mathcal{L}_{MSE}(y, G(x;\theta)) = \|y - G(x;\theta)\|_2^2 . \quad (1)$$

Given a training set $\{I_i\}_{i=1}^N$ where $I_i = \{x_i, y_i, \ell_i\}$, the generator network G is trained by minimizing $\mathcal{L}_{MSE}$. To be specific, the optimal $\hat{\theta}$ is obtained by solving the following optimization problem:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{MSE}(y_i, G(x_i;\theta)). \quad (2)$$

The ideal case is to have $y = G(x;\hat{\theta})$, *i.e.*, the trained generator network G works perfectly and outputs the true modality. However, in practice, minimizing the content loss function above can hardly achieve this, and the reason is clear in the probabilistic perspective. Minimizing $\mathcal{L}_{MSE}$ can be interpreted as maximizing the likelihood of $p_G(y|x;\theta)$ by assuming that $p_G(y|x)$ follows a Gaussian distribution. It is equivalent to minimizing the Kullback-Leibler (KL) divergence $D_{KL}(p_Y||p_G)$.

This training strategy suffers from the so-called "blurry" problem when p_Y follows a complex distribution, such as a multimodal distribution, which is usually the case in multi-modality missing data completion tasks. For example, in Figure 1, suppose the underlying distribution p_Y is a mixture of two Gaussian distributions with two equally probable modes g_1 and g_2. The optimized p_G will be a unimodal Gaussian with a single mode $(g_1 + g_2)/2$ by minimizing $D_{KL}(p_Y||p_G)$. Consequently, the modalities predicted by G tend to be blurry due to the average of two modes.

Another problem of the content loss function $\mathcal{L}_{MSE}$ is that it only minimizes element-wise differences between the predicted modalities and the true ones. It does not take any global similarity, like structural similarity between modalities, into consideration. To generate high-quality and informative modalities, we need to address the blurry problem and incorporate global information.

3.3 Adversarial Loss

In order to address the limitations suffered by the content loss $\mathcal{L}_{MSE}$, we propose to use the adversarial loss, based on generative adversarial networks (GANs) [10, 26]. In addition to the generator network G, a discriminator network D, parameterized by β, is employed. To be concrete, given the modality x and y of a subject, the discriminator $D((x, y);\beta)$ distinguishes whether the pair (x, y) is real or fake. Hence, D is a binary classification network. The corresponding training data is (x, y) from the subject of given dataset with label 1 and $(x, G(x;\theta))$ with label 0, where G is a generator network.

The objective function of adversarial learning in our model can be expressed as follows:

$$\min_{\theta} \max_{\beta} \mathbb{E}_{x\sim P_X, y\sim P_Y}[\log D(x, y)] + \mathbb{E}_{x\sim P_X}[1 - D(x, G(x))], \quad (3)$$

where the generator G is parameterized by θ and the discriminator D is parameterized by β. We optimize the adversarial loss based on minimax game theory. In this process, we first train the discriminator D to distinguish the true modality pairs from predicted modality pairs. Hence, the discriminator is a binary classifier. We give the true modality pair (x, y) a label of 1 and the predicted modality pair $(x, G(x))$ a label of 0. We minimize the following cross-entropy loss to train the classifier:

$$\mathcal{L}_{CE}(\hat{c}, c) = -\sum_{i=1}^{k} \mathbf{1}\{c = i\} \log \hat{c}_i, \quad (4)$$

where k is the number of classes, $c \in \mathbb{R}$ is the true label, $\hat{c} \in \mathbb{R}^{k\times 1}$, $\hat{c}_i$ is the probability that the sample belongs to category i, and $\mathbf{1}\{\cdot\}$ is the indicator function.

In this work, the discriminator is a binary classifier. Hence, $k = 2$ in Eq. (4). The loss function of the discriminator D can be expressed as:

$$\mathcal{L}_D(x, y) = \mathcal{L}_{CE}(D(x, y), 1) + \mathcal{L}_{CE}(D(x, G(x)), 0). \quad (5)$$

Given the same dataset in Section 3.2, the optimal β can be obtained by solving the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_D(x_i, y_i). \quad (6)$$

The generator is optimized to estimate the missing data so that the generated data is hard to be distinguished from the true data by the discriminator. Therefore, we train the generator with the following objective function $\mathcal{L}_G$ by fixing the parameter β in discriminator:

$$\mathcal{L}_G(x) = \mathcal{L}_{CE}(D(x, G(x)), 1). \quad (7)$$

Given the same dataset in Section 3.2, the optimal θ can be obtained by solving the following optimization problem:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_G(x_i). \quad (8)$$

By optimizing the generator and the discriminator iteratively, the generator can generate the missing modality data that is hard to be distinguished from the true data by the discriminator.

To see why the adversarial loss can alleviate the blurry effect caused by optimizing $\mathcal{L}_{MSE}$, we again use the probabilistic perspective. Given the same example in Section 3.2, the optimized generator network may not predict an averaged-mode data $(g_1 + g_2)/2$, since this data can be distinguished from the true data by the discriminator network. To fool the discriminator, the generator must predict the modality which is close to the true distribution. Therefore, employing the adversarial loss can alleviate the blurry effect of $\mathcal{L}_{MSE}$.

In addition, the discriminator takes the modality pair as input and distinguishes whether the pair is true or predicted. It does not consider the distance between each position in the true and predicted modalities. Therefore, optimizing the adversarial loss also encourages the predicted modality to be close to the true modality in a global view.

3.4 Classification Loss

The category label is of great importance in multi-modality missing data completion tasks. The input data has different categories. For subjects from different categories, the relationship between the input modality and the missing modality can be different. Therefore, it is necessary to take category label into consideration for the missing data completion task. The challenge is that the category label is not available when completing the missing modality on test data. To overcome this problem, we propose to employ an auxiliary classification loss in the discriminator to distinguish the different categories of inputs.

In the proposed model, the discriminator produces not only the real/fake probability distribution but also the category probability distribution of the input pairs. Therefore, the classification loss consists of two parts; those are, the cross entropy loss for true pairs and predicted pairs. The definition of $\mathcal{L}_D$ has been discussed in Section 3.3. The classification loss function $\mathcal{L}_{CLS}$ can be described as follows:

$$\mathcal{L}_{CLS}(x, y, l) = \mathcal{L}_{CE}(D(x, y), \ell) + \mathcal{L}_{CE}(D(x, G(x)), \ell). \quad (9)$$

The discriminator is trained to minimize $\mathcal{L}_D + \mathcal{L}_{CLS}$ and the generator is trained to minimize $\mathcal{L}_G + \mathcal{L}_{CLS}$. The predicted modality is not only close to the true modality in order to fool the discriminator but also takes the category information into consideration to optimize $\mathcal{L}_{CLS}$. Given the same dataset in Section 3.2, the optimal β can be obtained by solving the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_D(x_i, y_i) + \mathcal{L}_{CLS}(x_i, y_i, \ell_i)]. \quad (10)$$

The optimal θ can be obtained by solving the following optimization problem:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} [\mathcal{L}_G(x_i) + \mathcal{L}_{CLS}(x_i, y_i, \ell_i))]. \quad (11)$$

Another reason why we employ an auxiliary classification loss in the discriminator is that it can make the training procedure stable. The unstable training procedure makes adversarial networks difficult to train. It has been shown in prior work [25] that the category label information can make the training stable, leading to improved quality of the predicted sample.

3.5 The Proposed Optimization Problem

Training the generator with $\mathcal{L}_G + \mathcal{L}_{CLS}$ encourages the generator to produce data which confuse the discriminator. The adversarial loss distinguishes the predicted and true data in a global perspective. The auxiliary classifier can make the training procedure stable. However, if we only employ the adversarial loss and classification loss to optimize the generator, the generator can generate data that has a similar contour to the true modality. This predicted data can confuse the discriminator, but it loses many details. To overcome this problem, we combine the $\mathcal{L}_{MSE}$, $\mathcal{L}_D$ and $\mathcal{L}_{CLS}$ in our optimization. $\mathcal{L}_{MSE}$ encourages the learning of detailed information for completing the missing modality. The adversarial loss is employed to alleviate the blurry effect of $\mathcal{L}_{MSE}$ loss and improve the quality of predicted data. The classification loss is used to make the training procedure stable and take category label into consideration to improve the completion performance. Therefore, the overall loss function can be described as follows:

$$\mathcal{L} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_G\mathcal{L}_G + \lambda_{CLS}\mathcal{L}_{CLS}, \quad (12)$$

where λ_{MSE},λ_G,λ_{CLS} are tradeoff parameters for each loss. The algorithm for optimizing the problem in Eq. (12) is given in Algorithm (1).

Algorithm 1: Training the proposed deep adversarial networks for missing modality completion.

Set the batch size m, learning rates ρ_D and ρ_G, and weights λ_G,λ_{MSE}, λ_D, λ_{CLS}.
while *not converged* **do**
  Update the discriminator D**:**
  • Get m pairs of modality and category label $(x_1, y_1, l_1), \cdots, (x_m, y_m, l_m)$
  • Update the discriminator by:
  $\beta = \beta - \rho_D \nabla_\beta \frac{1}{m} \sum_{i=1}^{m} [\lambda_D \mathcal{L}_D(x_i, y_i) + \lambda_{CLS}\mathcal{L}_{CLS}(x_i, y_i, l_i)]$
  Update the generator G**:**
  • Get m new data samples $(x_1, y_1, l_1), \cdots, (x_m, y_m, l_m)$
  • Update the generator by:
  $\theta = \theta - \rho_G \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} [\lambda_{MSE}\mathcal{L}_{MSE}(y_i, G(x_i)) + \lambda_G \mathcal{L}_G(x_i) + \lambda_{CLS}\mathcal{L}_{CLS}(x_i, y_i, l_i)]$

3.6 Brain Data Completion

In this work, we focus on the brain multi-modality [29] missing data completion problem as illustrated in Figure 3. Specifically, x is the MRI modality which is available for all subjects, and y is the PET modality which is missing for some subjects. In addition, we are also given the category labels in the dataset, which refer to the prodromal stage including Normal controls (NC), Alzheimer's disease (AD), sMCI (the patient's symptom is stable and will not progress to AD in 18 months) , pMCI (the patient will progress to AD in 18 months).

In order to predict the PET modality from MRI modality, we need to construct a generator which can capture the relationship between the MRI modality and PET modality. Since both modalities are 3D data and the PET modality has the same spatial size as the MRI modality, we construct a 3D encoder-decoder network with skip connection as the generator [2, 4, 27]. The encoder network consists of 3D convolutional layers [15]. The decoder network consists of 3D convolutional layers and 3D deconvolutional layers. The 3D convolutional layers are used to extract features and predict the values of voxels in the PET modality. We employ the 3D convolutional layers with stride $2\times2\times2$ in the network to reduce the size of feature maps and increase the receptive fields of output voxels. The encoder-decoder architecture can extract features from the MRI modality with different scales. Because the MRI modality and the PET modality have the same size, 3D deconvolutional layers [9] are used to restore the spatial information. We add skip connections [11, 12] in the network between the corresponding encoder and decoder layers to allow the information to transmit directly in each level of the network.

The whole pipeline of brain modality completion is illustrated in Figure 2. The 3D encoder-decoder network takes the MRI modality as input and generates the PET modality. The predicted PET modality is concatenated with the MRI modality and used as an input pair for the discriminator network. The true modality pair is also fed as input to the discriminator network. The discriminator network is trained with the combined loss of $\mathcal{L}_D$ and $\mathcal{L}_{CLS}$. The generator network is trained with the combined loss of $\mathcal{L}_{MSE}$, $\mathcal{L}_G$ and $\mathcal{L}_{CLS}$. When completing the PET modality, only the generator network is required to take the MRI modality as input.

4 EXPERIMENTS

In this section, we evaluate our proposed model on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [32] and compare the quality of predicted data using different loss functions. Our code is publicly available[1].

4.1 Data Preprocessing

In this work, we use the data for 398 subjects in the ADNI dataset. Each subject has an MRI modality and a corresponding PET modality [1, 30]. These subjects cover four different prodromal stages (NC, sMCI, pMCI, AD). There are 93 AD subjects, 101 NC subjects, 76 pMCI subjects, and 128 sMCI subjects in the dataset.

For each subject in the dataset, we correct the intensity inhomogeneity and remove the cerebellum in the T1-weighted MRI modality. The MRI modality is segmented into gray matter, white matter and cerebrospinal fluid. In the experiments, we use the gray matter tissue density maps as the input modality and predict the output PET modality. The PET modality obtained from the ADNI dataset is also rigidly aligned to the corresponding MRI modality. We employ a unit standard deviation Gaussian kernel to smooth both the MRI and PET modalities in order to improve the signal-to-noise ratio. We downsampled both the input and output modalities to $64\times64\times64\times1$ voxels to reduce the computational cost.

[1] https://github.com/divelab/completion/

4.2 Experimental Setup

Table 1: The list of hyperparameter values for the generator. BN represents batch normalization; FMs represents the number of feature maps. The input size is $64\times64\times64\times1$. The batch size is equal to 5. The α, β_1, β_2 of Adam are set to 0.001, 0.5 and 0.999, respectively.

Type	Kernel	Stride	FMs	BN	Nonlinearity
3D-Conv	3	1	16	√	ReLU
3D-Conv	3	2	16	√	ReLU
3D-Conv	3	1	32	√	ReLU
3D-Conv	3	2	32	√	ReLU
3D-Conv	3	1	64	√	ReLU
3D-Conv	3	2	64	√	ReLU
3D-Conv	3	1	128	√	ReLU
3D-Conv	3	2	128	√	ReLU
3D-Conv	3	1	256	√	ReLU
3D-Conv	3	1	256	√	ReLU
3D-Deconv	3	2	128	√	ReLU
3D-Conv	3	1	128	√	ReLU
3D-Deconv	3	2	64	√	ReLU
3D-Conv	3	1	64	√	ReLU
3D-Deconv	3	2	32	√	ReLU
3D-Conv	3	1	32	√	ReLU
3D-Deconv	3	2	16	√	ReLU
3D-Conv	3	1	1	×	None

Table 2: The list of hyperparameter values for the discriminator. BN represents batch normalization; FMs represents the number of feature maps. The input size is $64\times64\times64\times2$, and the batch size is equal to 5. The slope of leaky ReLU equals to 0.2. The α, β_1, β_2 of Adam are set to 0.0002, 0.5 and 0.999, respectively.

Type	Kernel	Stride	FMs	BN	Dropout	Nonlinearity
3D-Conv	3	1	16	√	0.5	Leaky ReLU
3D-Conv	3	2	32	√	0.5	Leaky ReLU
3D-Conv	3	1	64	√	0.5	Leaky ReLU
3D-Conv	3	2	128	√	0.5	Leaky ReLU
Linear	N/A	N/A	5	×	0.0	Soft-Sigmoid

Based on the descriptions in Section 3, we construct a 3D encoder-decoder network as a generator for PET modality completion task.

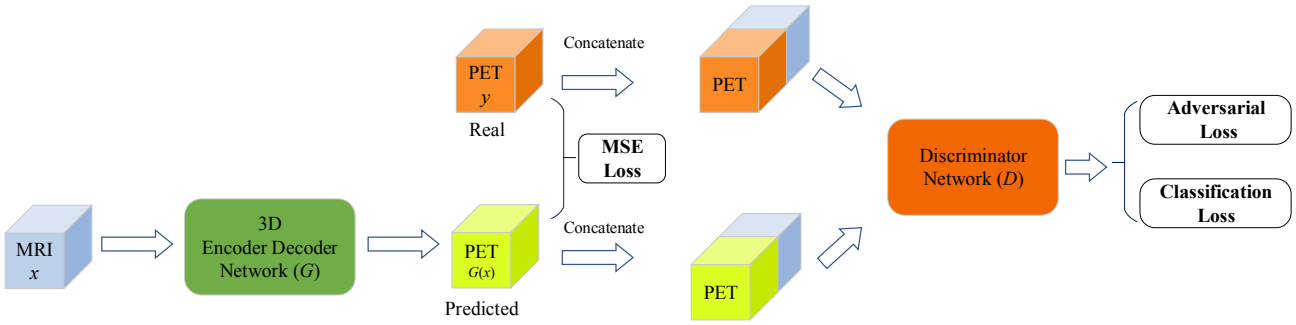

Figure 2: Illustration of our proposed model for the brain data completion task. In the model, the 3D encoder-decoder generator network takes the MRI modality as the input and generates the PET modality. The discriminator takes a subject with both modalities as input and predicts whether the two modalities match as well as the category of the subject.

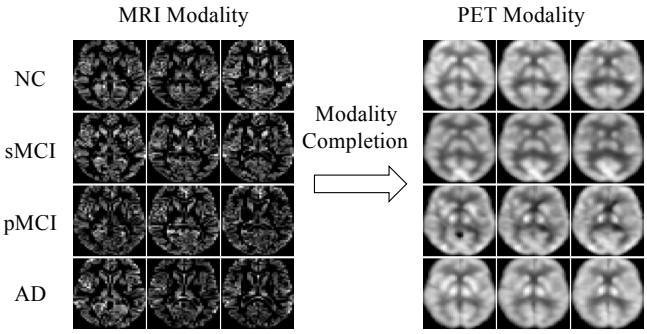

Figure 3: Illustration of the multi-modality missing data completion problem in which the MRI modality is given as the input and the PET modality is predicted as the output.

The network architecture is shown in Table 1. The depth of the generator network is 5. Each layer contains a downsampling block and a corresponding upsampling block. In the downsampling block, we employ two 3D convolutional layers to extract features. To facilitate the training, we use batch normalization [13] after each convolutional layer. The stride of the second convolutional layer in the downsampling block is set to $2 \times 2 \times 2$ to increase the receptive fields of voxels in output. The number of output feature maps is 16 in the first downsampling block. For the following blocks, the number of output feature maps is twice of the number in the previous block. In terms of the upsampling block, we employ 3D deconvolutional layer to restore the spatial size. We use skip connections between the upsampling block and the corresponding downsampling block to directly transmit information. The concatenated feature maps are fed into a 3D convolutional layer. The learning rate [16] of the generator network is set to 1e-3, and the batch size is set to 5.

We employ a 3D convolutional network for the discriminator. The architecture of the network is shown in Table 2. The network consists of four 3D convolutional layers and one fully connected layer. Since the adversarial loss is difficult to train, we employ several techniques in training. Specifically, we use the batch normalization layer after each convolutional layer. Instead of using max-pooling layers, we employ 3D convolutional layers with stride $2 \times 2 \times 2$ to implement downsampling. We use dropout layers with probability 0.5 after each batch normalization layer. In tradition adversarial networks, the number of output nodes for fully connected layers is set to one. To make the training procedure stable, we add a classification loss in the discriminator network. Since the PET modality covers four different categories, the number of output nodes for the fully connected layer is set to five. One node represents whether the pair of MRI and PET modalities is real or fake. The remaining four nodes represent the probabilities of each category for the pair of modalities. The learning rate of the discriminator is set to 2e-4.

Table 3: Comparison of similarity and sharpness measures by the three models on test dataset.

	Similarity		Sharpness
	PSNR	SSIM	
$\mathcal{L}_{MSE}$	34.68	0.8600	23.83
$\mathcal{L}_{MSE} + \mathcal{L}_G$	34.50	0.9836	24.09
$\mathcal{L}_{MSE} + \mathcal{L}_G + \mathcal{L}_{CLS}$	**34.90**	**0.9854**	**24.26**

In this work, we focus on evaluating our proposed method for missing data completion. The ADNI dataset contains 398 subjects and each subject has a pair of MRI and PET modalities. We randomly select 200 subjects in the dataset as the training set to train the model. We evaluate the performance of the generators with different losses on the remaining 198 subjects.

4.3 Results and Analysis

In the experiments, we train the 3D encoder-decoder network using different loss functions and evaluate the predicted modality on the test dataset. To evaluate the quality of predicted modality and the similarity between the predicted modality and the true modality, we employ different quantitative evaluation metrics. The results are summarized in Table 3.

We employ the peak signal-to-noise ratio (PSNR) [28] to measure the quality of the predicted modality. PSNR is generally used to measure the quality of predicted data in video prediction tasks. The

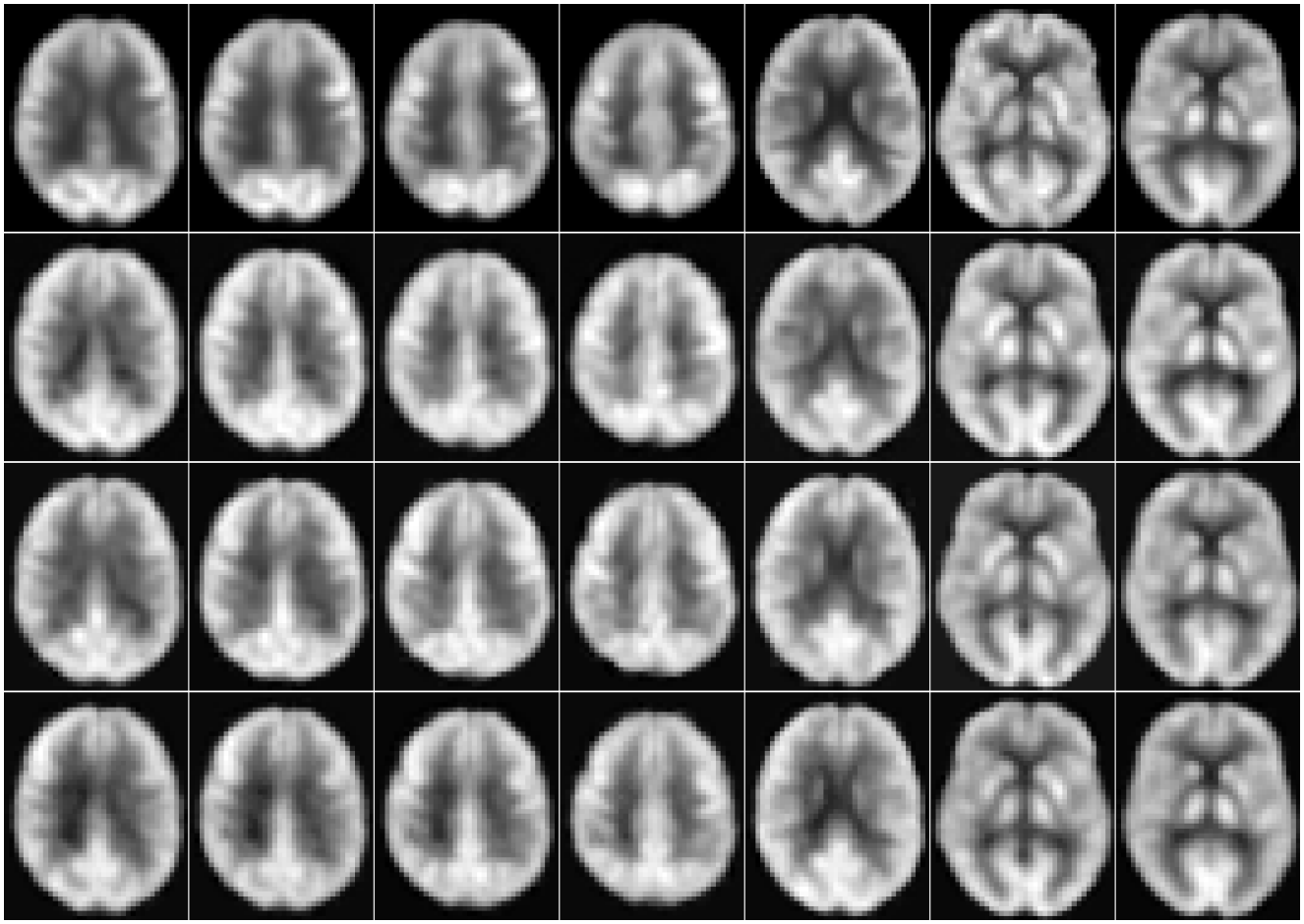

Figure 4: Some sample data completion results. The first row is the ground truth; the second row is the results predicted by $\mathcal{L}_{MSE}$; the third row is the results predicted by $\mathcal{L}_{MSE}$+$\mathcal{L}_G$. The last row is the results predicted by $\mathcal{L}_{MSE}$+$\mathcal{L}_G$+$\mathcal{L}_{CLS}$.

definition of PSNR between the predicted modality $\hat{y}$ and the true modality y is defined as follows:

$$\text{PSNR}(\hat{y}, y) = 10 \log_{10} \frac{\max_{\hat{y}}^2}{\mathcal{L}_{MSE}(\hat{y}, y)}, \tag{13}$$

where $\max_{\hat{y}}$ represents the maximum possible voxel value of the modality. Larger PSNR values imply larger similarity between the predicted modality and the true modality.

We can observe from the results in Table 3 that the generator using the combination of $\mathcal{L}_{MSE}$, $\mathcal{L}_G$ and $\mathcal{L}_{CLS}$ achieves the highest PSNR value. The definition of PSNR value is given based on the MSE value. The generator that only uses $\mathcal{L}_{MSE}$ as a loss function obtains better PSNR value than the generator using $\mathcal{L}_{MSE}$ and $\mathcal{L}_G$. Compared with the generator using $\mathcal{L}_{MSE}$, the model using $\mathcal{L}_{MSE}$ and $\mathcal{L}_G$ has to balance $\mathcal{L}_{MSE}$ and $\mathcal{L}_G$. Therefore, the PSNR value of the generator with $\mathcal{L}_{MSE}$ is better than the model with the combination of $\mathcal{L}_G$ and $\mathcal{L}_{MSE}$. After using $\mathcal{L}_{CLS}$, the generator achieves the best PSNR value on the test dataset. This also demonstrates that our model has a better generalization ability.

We also use the structural similarity index measure (SSIM) [31] to measure the similarity between the predicted and the true modalities. Different from the MSE and PSNR measurements, SSIM does not measure the absolute difference. It focuses on the difference in terms of structural information. The SSIM score is calculated using the following equation:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{14}$$

where μ_x is the mean of x, μ_y is the mean of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, c_1 and c_2 are two constant variables. The value of SSIM score ranges between -1 and 1.

The PSNR score measures the voxel-wise absolute error between the two modalities. The SSIM score evaluates the similarity between two modalities based on the structure information. The model using the combination of $\mathcal{L}_{MSE}$ and $\mathcal{L}_G$ obtains a lower PSNR value compared with the generator employing $\mathcal{L}_{MSE}$. But it achieves significant improvement on the SSIM score. Since the discriminator distinguishes the real/fake modality from a global perspective, the

Table 4: Comparison of classification accuracy by the three models on test dataset.

	Multi-class classification	AD - NC	AD - pMCI	AD - sMCI	NC - pMCI	NC - sMCI	pMCI - sMCI
MRI	37.94	82.29	**51.19**	73.83	72.72	**58.65**	55.57
True PET	38.97	**83.33**	50.00	**74.76**	**75.00**	58.55	62.62
$\mathcal{L}_{MSE}$	38.46	73.95	46.42	69.15	65.90	51.35	63.63
$\mathcal{L}_{MSE} + \mathcal{L}_G$	39.48	76.04	**50.00**	65.42	71.59	55.85	59.59
$\mathcal{L}_{MSE} + \mathcal{L}_G + \mathcal{L}_{CLS}$	**40.00**	72.91	48.8	**74.76**	64.77	**59.45**	**65.65**

modality predicted using the combination loss has better performance on maintaining structure information and thus achieves a higher SSIM score. We can observe from the results that employing $\mathcal{L}_{CLS}$ in the discriminator improves the performance in terms of both PSNR and SSIM scores. This demonstrates that the category information is important for the PET modality completion task. Our model can predict the missing PET modality without providing the category label as the input.

One key challenge in modality completion tasks is to produce sharp images. To measure the sharpness of predicted modality and show the effects of adversarial loss. We employ the sharpness score proposed in [23] to measure the sharpness between the predicted modality and the true modality. The definition of sharpness is described as follows:

$$\text{Sharp}(x, y) = 10 \log_{10} \frac{\max_Y^2}{\frac{1}{N}\left(\sum_i \sum_j \sum_k |(\nabla_i x + \nabla_j x + \nabla_k x) - (\nabla_i y + \nabla_j y + \nabla_k y)|\right)},$$

where

$$\begin{aligned} \nabla_i x &= |x_{i,j,k} - x_{i-1,j,k}|, \\ \nabla_j x &= |x_{i,j,k} - x_{i,j-1,k}|, \\ \nabla_k x &= |x_{i,j,k} - x_{i,j,k-1}|. \end{aligned}$$

The sharpness is calculated based on the gradient between the voxels in the original modality. A higher score is obtained if the gradient in the predicted modality is close to the gradient in the original modality.

We can observe from the results that our proposed method outperforms the generator using $\mathcal{L}_{MSE}$. $\mathcal{L}_{MSE}$ is employed based on the assumption that the data follow a single Gaussian distribution. When the subjects in dataset follow a more complex distribution, the generator using $\mathcal{L}_{MSE}$ predicts blurry modality. The sharpness values in the table also verify this observation. The generator model using $\mathcal{L}_{MSE}$ obtains the lowest sharpness value. When we employ an additional adversarial loss in the model, the effect of $\mathcal{L}_{MSE}$ is alleviated and the PET modality predicted using the combined loss is sharper. This demonstrates that the discriminator can successfully distinguish the averaged modality from the true modality. By introducing the adversarial loss, the model can produce sharper images. The adversarial loss is difficult to train. We employ a classification loss in discriminator network to make the training procedure stable. Thus, the generator with the combination of three losses achieves the best sharpness value.

In Figure 4, we provide the true PET slices and predicted PET slices using different loss functions. We can observe from the results that the images obtained from the generator using the combination of three losses is closer to the true PET modality as compared with the results of other models. In the true PET modality, there exists a clear contour in the image. The images generated by the model with the MSE loss are blurry and the contour of the components is also not clear. By introducing the adversarial and classification loss, the effect of $\mathcal{L}_{MSE}$ is alleviated.

We also compare the classification results of the true PET modality and the predicted PET modalities using different losses. If the predicted modality is close to the corresponding true PET modality, the category label should be the same. Therefore, we employ a logistic regression model as a classifier to evaluate the predicted modality using different losses.

We train a logistic regression classifier based on the PET modality in the training dataset and test the classification accuracy based on the predicted modality using different losses in the test dataset. The results of classification accuracy are shown in Table 4. The PET modality covers four different categories. In Table 4, we provide both multi-class classification accuracy and binary classification accuracy. We can observe from the results that the model with three losses achieves higher multi-class classification accuracy than the other models, including that of the true PET modality. The development of Alzheimer's disease includes four stages (NC, sMCI, pMCI, AD). We extract subjects from each category and construct a binary classifier. The challenge in the diagnosis of Alzheimer's disease is to distinguish the neighboring stages. We can also observe from the results that the classification accuracy of neighboring stages is lower than that of other stages. For example, the classification accuracies of NC/sMCI, sMCI/pMCI, pMCI/AD are lower than that of others. The classification accuracies of our model for AD/NC and NC/pMCI tasks are lower than the model with $\mathcal{L}_{MSE}$. However, these two tasks can be successfully completed by only using the MRI modality. In terms of neighboring stages classification tasks, our predicted data outperforms other modalities, including that of the true PET modality. This demonstrates that our predicted PET modality can be used to improve the accuracy of disease diagnosis.

5 CONCLUSION AND DISCUSSION

In this work, we propose a deep learning model for completing the missing modality and apply it to Alzheimer's disease diagnosis. Our model takes the MRI modality as input and predicts the corresponding PET modality, which can improve the accuracy of Alzheimer's disease diagnosis. Our model employs a 3D encoder-decoder network to capture the relationship between the MRI modality and PET modality. To alleviate the blurry effect of $\mathcal{L}_{MSE}$ and generate high quality data, we employ an adversarial loss in our model. Since the category label information is of key importance in data

completion, we add an additional classification loss in the discriminator. In this way, we can complete the missing modality without the category label information as an input. The classification loss can also make the training procedure stable. Experiment results on the ADNI dataset demonstrate that our predicted PET modality achieves higher quality both in similarity and sharpness compared to the predicted modality with $\mathcal{L}_{MSE}$. In addition, our predicted PET modality outperforms other predicted modalities in classification tasks. This also demonstrates our model can improve the accuracy of disease diagnosis.

In this work, we focus on completing one modality data based on data from another modality. Our method can be applied to complete multiple modalities simultaneously based on data from another set of modalities. For example, another application in disease diagnosis is to predict high-dose PET data based on MRI and low-dose PET data [1, 30], as high-dose PET data are not available for many patients. We plan to explore those applications in the future.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants DBI-1641223, IIS-1615035, IIS-1633359, Defense Advanced Research Projects Agency grant N66001-17-2-4031, and National Institutes of Health grant EB008374.

REFERENCES

[1] Le An, Pei Zhang, Ehsan Adeli, Yan Wang, Guangkai Ma, Feng Shi, David S Lalush, Weili Lin, and Dinggang Shen. 2016. Multi-level canonical correlation analysis for standard-dose PET image estimation. *IEEE Transactions on Image Processing* 25, 7 (2016), 3303–3315.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 424–432.

[5] Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*. 1486–1494.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2016. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2016), 295–307.

[7] Hao Dong, Paarth Neekhara, Chao Wu, and Yike Guo. 2017. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676* (2017).

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. 2015. Learning to generate chairs with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 1538–1546.

[9] Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. 2017. Pixel Deconvolutional Networks. *arXiv preprint arXiv:1705.06820* (2017).

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 1. 3.

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017).

[15] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 221–231.

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

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint* (2016).

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

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[23] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).

[24] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[25] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2016. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585* (2016).

[26] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

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

[28] SC Strother, ME Casey, and EJ Hoffman. 1990. Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts. *Ieee transactions on nuclear science* 37, 2 (1990), 783–788.

[29] Qi Wang, Mengying Sun, Liang Zhan, Paul Thompson, Shuiwang Ji, and Jiayu Zhou. 2017. Multi-Modality Disease Modeling via Collective Deep Matrix Factorization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1155–1164.

[30] Yan Wang, Guangkai Ma, Le An, Feng Shi, Pei Zhang, David S Lalush, Xi Wu, Yifei Pu, Jiliu Zhou, and Dinggang Shen. 2017. Semisupervised tripled dictionary learning for standard-dose PET image prediction using low-dose PET and multimodal MRI. *IEEE Transactions on Biomedical Engineering* 64, 3 (2017), 569–579.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[32] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. 2013. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & dementia: the journal of the Alzheimer's Association* 9, 5 (2013), e111–e194.

[33] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. 2013. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 185–193.

[34] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. 2014. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage* 102, 0 (2014), 192–206.

[35] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. 2016. Multimodal deep learning for cervical dysplasia diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 115–123.

[36] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision*. 5907–5915.

[37] Shaoting Zhang and Dimitris Metaxas. 2016. Large-Scale medical image analytics: Recent methodologies, applications and Future directions. *Medical Image Analysis* 33 (2016), 98–101.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).