
Article:
Chen, Z., Wang, M., Nan, F. et al. (6 more authors) (2025) Joint image synthesis and fusion with converted features for Alzheimer's disease diagnosis. Engineering Applications of Artificial Intelligence, 156 (Part B). 111102. ISSN 0952-1976

https://doi.org/10.1016/j.engappai.2025.111102

Joint Image Synthesis and Fusion with Converted Features for Alzheimer's Disease Diagnosis

Zhaodong Chen[a], Mingxia Wang[a], Fengtao Nan[c], Yun Yang[a], Shunbao Li[b], Menghui Zhou[b,*], Jun Qi[d], Hanwen Wang[a] and Po Yang[b,*]

[a]National Pilot School of Software, Yunnan University, Kunming, 650500, China
[b]Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, U.K
[c]Northwest A&F University, College of information engineering, Northwest A&F University, Xianyang, 712199, China
[d]Department of Computing, School of Advanced Technology, Xian JiaoTong-Liverpool University, Suzhou, 215028, China

ARTICLE INFO

Keywords:
Cross-modality synthesis
Generative adversarial networks
Multi-modal fusion
Joint optimization

ABSTRACT

The effectiveness of complete multi-modal neuroimaging data in the diagnosis of Alzheimers disease has been extensively demonstrated and applied. Dealing with incomplete modalities poses a common challenge in multi-modal neuroimaging diagnosis. The mainstream approaches aim to synthesise missing neuroimaging data in order to make full use of all available samples. However, these methods treat image synthesis and disease diagnosis as two independent tasks, overlooking the potential feature of cross-modality image synthesis for downstream tasks. To this end, we propose the Joint Image Synthesis and Classification Learning method to jointly optimize image synthesis and disease diagnosis using incomplete neuroimaging modalities. Our approach comprises a submodule for synthesising missing neuroimaging data and a decision fusion submodule that integrates features from different modalities and the high-level/converted features generated during synthesis. Experimental results demonstrate that our joint optimization approach outperforms conventional two-stage methods. Our method is capable of handling arbitrary neuroimaging modality missing scenarios and achieves state-of-the-art performance in both Alzheimers Disease identification and mild cognitive impairment conversion classification tasks. Finally, we further explored the importance of different converted features. This highlights the effectiveness of our approach in addressing the challenges of Alzheimers Disease diagnosis and provides insights for future research in multi-modal medical image analysis.

1. Introduction

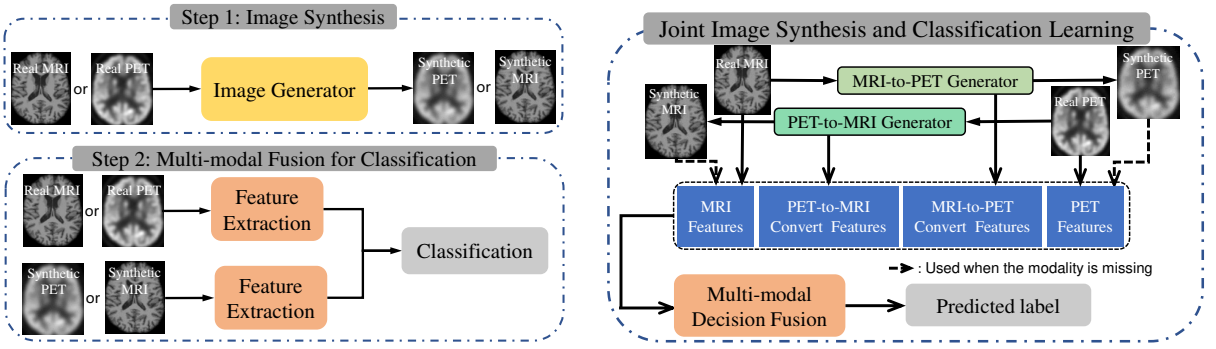
Alzheimer's disease (AD) is the most common neurodegenerative disorder [1, 2, 3, 4]. It is characterized by symptoms including memory loss, cognitive function deterioration, as well as language and behavioral issues. These symptoms significantly impact patients' daily lives, causing distress and challenges.

Multi-modal neuroimaging data, such as structural MRI and Fluorodeoxyglucose PET, have demonstrated their effectiveness in enhancing the diagnostic performance for AD [5, 6, 7, 8]. In practice, the availability of specific modalities for each subject is often limited due to various challenges, including high costs, lengthy acquisition times, image corruption, and privacy concerns. For instance, in the Alzheimer's Disease Neuroimaging Initiative (ADNI-1) dataset [9], only 360 out of 709 subjects have fully paired PET and MRI data.

When addressing the issue of missing modalities, two predominant machine learning methods are commonly employed. Traditional methods often discard incomplete samples from modalities [10, 5, 11]. This practice can result in a high-dimensional small sample problem and a consequent reduction in diagnostic performance. Additionally, researchers have introduced machine learning-based data imputation methods to overcome this limitation, aiming to estimate missing data based on complete subject features. For example, Marlin et al. [12] covers different types of missing data and their causes, as well as various approaches to handle missing data, including deletion and imputation methods. Thung et al. [13] proposes a method that combines matrix shrinkage and completion techniques to diagnose neurodegenerative diseases using incomplete multi-modality data. Notably, the aforementioned methodologies rely

chenzhaodong2@163.com (Z. Chen); mingxiacn@163.com (M. Wang); fengtao.nan@nwsuaf.edu.cn (F. Nan); yunyang@ynu.edu.cn (Y. Yang); shunbao.li@sheffield.ac.uk (S. Li); mzhou47@sheffield.ac.uk (M. Zhou); jun.qi@xjtlu.edu.cn (J. Qi); 15574676989@163.com (H. Wang); poyangcn@gmail.com (P. Yang)
ORCID(s): 0009-0008-1737-817X (Z. Chen); 0000-0002-8553-7127 (P. Yang)

on manually extracted features, potentially failing to encompass all disease-relevant characteristics and thereby constraining diagnostic accuracy.



(a) Two-stage image synthesis and multi-modal classification learning (b) Our proposed method for Joint Image Synthesis and Classification Learning

Figure 1: Comparison of the conventional method and the method in this paper.

In recent years, with the advancement of Generative Adversarial Networks (GAN), several cross-modality synthesis algorithms have emerged. Huang et al. [14] proposes a method for cross-modality image synthesis in MRI using joint dictionary learning, which incorporates a geometric constraint to improve the synthesis process. Additionally, Maspero et al. [15] focuses on the evaluation of a fast synthetic-CT generation method using a pix2pix [16] method. Moreover, Pan et al. [17] utilizes a Cycle-Consistent GAN (CycGAN) to synthesise missing PET images from MRI data, leveraging the cycle-consistency constraint to learn the mapping between MRI and PET images and generate synthetic PET images.

The above strategies have some limitations. Firstly, they primarily focus on cross-modality synthesis without thoroughly evaluating the effectiveness of the synthesised images for downstream tasks. Secondly, these approaches are tailored to address particular missing modalities but lack the capacity to handle arbitrary modality gaps within the data. This implies that individual image generators must be trained for each modality. To mitigate these drawbacks, Hu et al. [18] proposed a method that utilizes two GAN: one for MRI-to-PET synthesis and another for PET-to-MRI synthesis. These GANs consist of generator and discriminator networks trained in an adversarial manner to generate realistic images and distinguish real from synthesised images. This approach still treats image synthesis and downstream classification tasks as two separate tasks.

In light of these considerations, we propose a unified framework called Joint Image Synthesis and Classification Learning (JISCL) to address the challenge of incomplete multi-modal neuroimaging data. The JISCL framework is trained and tested on incomplete modalities, as illustrated in Figure 1(b). It combines cross-modality synthesis and multi-modal fusion learning to improve diagnostic performance. The cross-modality synthesis component of the JISCL framework employs two Multi-Scale Generative Adversarial Network (MGAN) architectures. These architectures are responsible for transforming between the MRI and PET modalities and extracting converted features during the image generation process. These converted features play a crucial role in disease diagnosis. The multi-modal fusion component integrates private features extracted from the MRI and PET modalities with the public/converted features obtained from the image generation process. These features are fused at the decision level to enhance the classification task. For more detailed information on the network architectures employed in the JISCL framework, please refer to Section 3.2 of the paper.

To address the challenges of missing modalities and multi-modal data heterogeneity in Alzheimer's disease diagnosis, this paper makes the following key contributions:

- We propose the **JISCL framework**, which unifies cross-modality image synthesis and multi-modal fusion learning within a single pipeline, rather than treating them as separate tasks.
- Unlike existing cross-modality synthesis methods that primarily focus on generating images, our framework extracts transformed features during the synthesis process to directly enhance diagnostic performance.

Table 1
Statistical information on the four categories of study subjects.

Label		CN			AD			sMCI			pMCI		
Dataset	modality	COM	ICOM	ALL	COM	ICOM	ALL	COM	ICOM	ALL	COM	ICOM	ALL
ADNI-1	MRI	**93**	111	**213**	**84**	86	**183**	**116**	94	**222**	**67**	82	**157**
	PET		9			13			12			8	
ADNI-2	MRI	**245**	3	**293**	**136**	2	**148**	**243**	6	**255**	**85**	1	**86**
	PET		45			10			6			0	

COM: Paired MRI and PET samples. ICOM: Samples with only MRI or PET.
ALL: Total samples after synthesizing missing neuroimaging data.

- We design a multi-modal fusion mechanism that effectively integrates **private features** (from individual modalities) and **public/converted features** (from image synthesis) at the decision level, improving classification accuracy.
- The proposed method does not rely on pre-trained modality-specific generators but instead provides a **generalizable** approach to synthesizing missing modalities, making it adaptable to various incomplete neuroimaging datasets.
- We introduce a **joint optimization strategy** that simultaneously trains the synthesis and classification components, ensuring that the synthesized images are optimized for downstream diagnostic tasks rather than just visual realism.

The remainder of this paper is organized as follows. Section 1 introduces the motivation and the innovation points of the experiment. Section 2 presents content related to cross-modality neural image synthesis and joint optimization. Section 3 presents the data, pre-processing methods, and the proposed approach. In Section 4, we first compare the performance of the proposed approach with three other generative models in cross-modal neuroimaging synthesis. Then, we compare the performance of the two-stage method with the JISCL diagnostic task and further compare our method with previous research. In Section 5, we analyze the limitations of the current work.

2. Related Work

Cross-modality neural image synthesis has garnered significant attention in recent years due to its ability to generate missing neuroimaging modalities. Several approaches have been developed to improve the quality and reliability of synthetic neuroimages. For instance, Hu et al. [19] propose a cross-modality MRI image synthesis framework utilizing neural architecture search, which automatically identifies the optimal network structure for synthesizing heterogeneous MRI modalities. Wang et al. [20] introduce an auto-context model that integrates both local and global contextual information, employing a multi-modality GAN with adversarial training and a locality-adaptive strategy to generate high-quality PET images from MRI inputs. These methods have demonstrated substantial improvements in image synthesis; however, they primarily focus on generating visually plausible images without explicit optimization for downstream diagnostic tasks.

2.1. Cross-Modality Image Synthesis for Neuroimaging

A common approach to evaluating the effectiveness of synthesized neuroimaging data involves a two-stage training paradigm (as illustrated in Figure 1(a)). In the first stage, a generative model synthesizes the missing neuroimaging modality. In the second stage, the generated images are incorporated into the dataset to facilitate multi-modal classification. For example, Sajjad et al. [21] introduce a deep convolutional GAN-based framework that generates synthetic PET images to augment limited training data, thereby enhancing model performance. Similarly, Sikka et al. [22] utilize the BPGAN framework to synthesize PET images from MRI scans, aiming to capture disease-specific features for multi-modal AD diagnosis. Kao et al. [23] further explore cross-modality synthesis by aligning the representations of T1-MRI and FDG^{18}-PET images in a shared latent space, ensuring high-quality image translation.

While these methods improve image synthesis fidelity, they treat image generation and classification as independent tasks, potentially leading to suboptimal diagnostic performance. The lack of an end-to-end optimization mechanism means that the synthesized images are not explicitly optimized for downstream clinical tasks, limiting their applicability in real-world diagnostic settings.

2.2. Joint Optimization of Image Synthesis and Classification

Recent research efforts have sought to integrate image synthesis with multi-modal classification in a unified framework. Liu et al. [24] propose an approach that leverages MRI and PET features during the image synthesis process to enhance multi-modal classification. However, their method only incorporates transformed features extracted during synthesis, neglecting the original private features of each modality. Similarly, Pan et al. [25] introduce a Feature Consistency GAN for missing neuroimaging data generation while simultaneously optimizing classification tasks. Nevertheless, their approach does not fully exploit the complementary information between different modalities, which could further improve diagnostic performance. Kläser et al. [26] present an imitation learning method for PET/MRI attenuation correction, training a deep neural network to mimic expert-labeled attenuation maps, thus improving accuracy and robustness.

Despite these advancements, existing methods still face limitations. Many studies primarily focus on synthesizing missing PET scans while neglecting the equally critical problem of missing MRI scans. Moreover, separating cross-modality image synthesis from multi-modal classification tasks fails to maximize the synergy between these components. Our approach addresses these shortcomings by introducing an end-to-end Joint Image Synthesis and Classification Learning (JISCL) framework that optimally integrates the generation and classification processes. By leveraging transformed features as a bridge, our method ensures that synthesized images contribute directly to improved diagnostic accuracy, providing a more holistic solution for handling incomplete neuroimaging data.

3. Meterials and Methods

3.1. Materials and image pre-processing

We preprocessed two subsets of the ADNI [9] dataset, namely ADNI-1 and ADNI-2. The ADNI is a collaborative research project focused on biomarkers and early detection of AD. It collects data on clinical, neuroimaging, genetics, and biomarkers to advance our understanding of the disease and support clinical trials and treatment evaluation. ADNI drives research for improved diagnostics and therapies. The dataset consists of four categories of subjects: Alzheimer's Disease (AD), Cognitively Normal (CN), progressive Mild Cognitive Impairment (pMCI) which progresses to AD within 36 months after baseline, and stable Mild Cognitive Impairment (sMCI) which does not progress to AD within 36 months after baseline. After removing duplicate subjects from ADNI-1 and ADNI-2, the dataset is summarized in Table 1. It is evident that there is a significant amount of missing modalities in ADNI-1. Out of 775 subjects, only 360 have paired MRI and PET data, while 415 subjects have data for only one modality. In ADNI-2, out of 782 subjects, 73 subjects have data for only one modality. Completing the missing modalities would result in a significant increase in the number of samples for each category in ADNI-1. The CN class increases from 93 to 213, the AD class from 84 to 183, the sMCI class from 116 to 222, and the pMCI class from 67 to 157. A similar trend can also be observed in ADNI-2.

These findings emphasize the crucial significance of cross-modality neuroimaging synthesis. By facilitating the generation of missing modalities, we can significantly enhance the dataset, thereby improving the resilience and applicability of models trained on such data. The increased sample size enables more thorough analysis and enhances the precision of neuroimaging diagnosis and classification tasks. This advancement contributes to the progress of neuroimaging research and its practical applications.

During the data pre-processing stage, we initially performed skull stripping on the MRI scans using the FreeSurfer software [27]. Subsequently, the PET and MRI scans were linearly aligned in pairs. To ensure spatial consistency between the paired MRI and PET, we employed SPM12 [28] for affine mapping of the MRI and PET onto a common MNI template. Following the pre-processing steps, all MRI and PET scans were resized to match the size of the MNI template ($182 \times 218 \times 182$) with isotropic voxels of $1 \times 1 \times 1$ mm. During training, we further resized the images to $128 \times 128 \times 128$ with isotropic voxels of $1.42 \times 1.70 \times 1.42$ mm to facilitate network construction and reduce the model parameters.

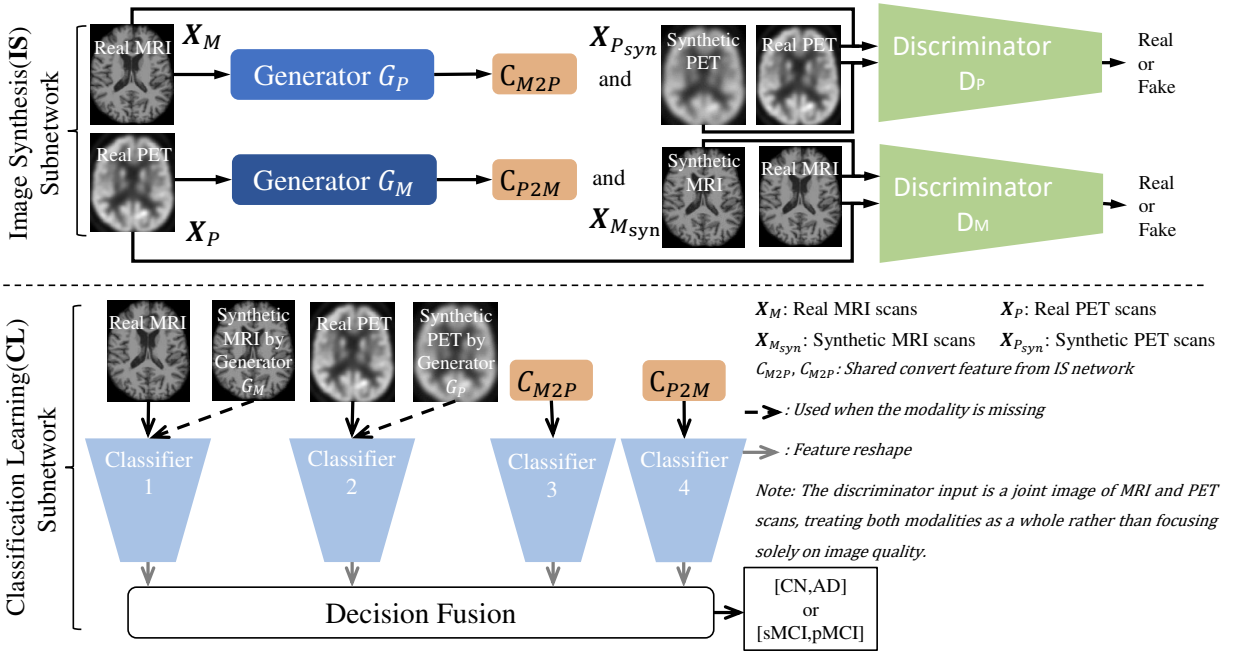


Figure 2: Illustration of the proposed Joint Image Synthesis and Classification Learning framework for AD prediction with an image synthesis sub-network and a classification learning sub-network.

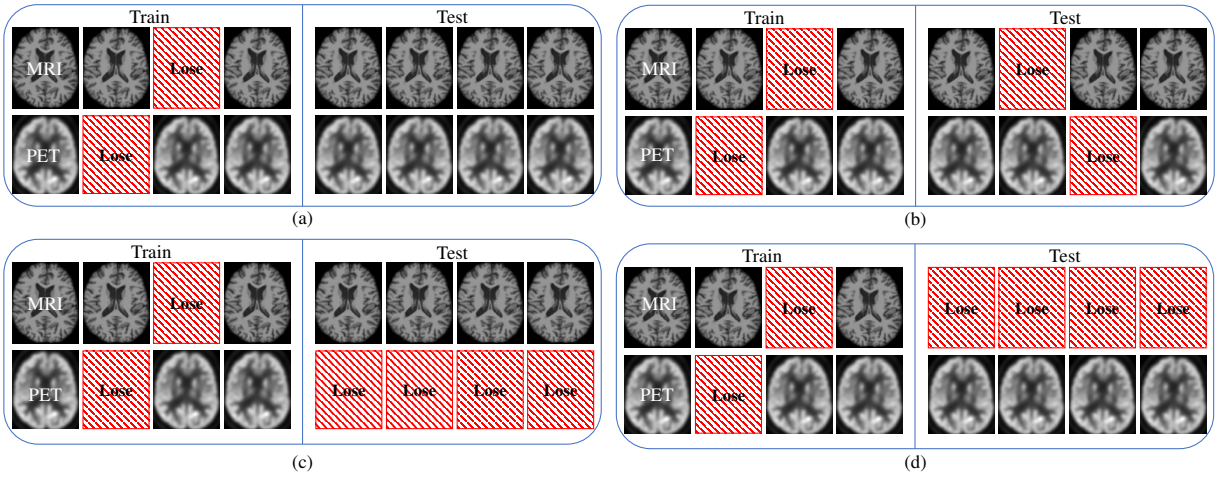


Figure 3: Illustration of different missing modality scenarios in neuroimaging data. The figure presents cases where MRI and PET images are either available or absent during training and testing phases.

We conducted two sets of comprehensive experiments using the ADNI-1 and ADNI-2 datasets. In the first set of experiments, our model was trained on the ADNI-2 dataset, and its performance was evaluated on the ADNI-1 dataset. Throughout this process, we synthesised missing MRI and PET images. In the second set of experiments, we trained our model on the ADNI-1 dataset and assessed its performance on the ADNI-2 dataset. Alongside synthesising missing MRI and PET images, we also performed brain disease diagnosis. To evaluate the generalization capability of our proposed JISCL model, we additionally tested the performance of the two-stage method using the dataset augmented with the synthesised modalities.

3.2. Method

The JISCL framework comprises two sub-networks: the Image Synthesis sub-network and the Classification Learning sub-network. These sub-networks share converted features, as depicted in Figure 2.

Image Synthesis Sub-network: This network is designed to handle the four different modality missing scenarios: Missing training modality only; Both training and testing modalities missing; Extreme testing modality missing (with only MRI or PET modality available). The image synthesis sub-network consists of two MGAN cross-modality image generators, which correspond to the cross-modality generation from MRI scans to PET scans and from PET scans to MRI scans. When both modalities are present in the sample, generator G_P generates PET scans from MRI scans, and generator G_M generates MRI scans from PET scans. When only one modality is present, for example, when the PET modality is missing, generator G_P first generates the missing PET scan, and then the generated PET scan serves as input for generator G_M to generate the MRI scan. The process is similar when the MRI modality is missing. When both modalities are complete, each generator performs the modality transformation task. The purpose of this approach is twofold: first, to enhance the capability of the image generators to produce more useful images when modality is missing; second, to generate transformation features for classification learning, providing more shared information. When only one modality is present, the two image generators form a structure similar to CycleGAN. Unlike CycleGAN, however, this model does not specify source and target modalities, and can build the network based on the actual situation, thus enhancing the effectiveness of image generation. In this case, in addition to synthesizing the missing modality, the source modality is also reconstructed. This approach allows us to investigate how much information from the source modality is preserved in the generated missing modality and enhances the generalization ability of the model by synthesizing missing images. Based on this strategy, this paper presents a clever solution to the incomplete modality problem. It effectively utilizes the complete modality to address the issue of incomplete modality and improves the effectiveness of the synthesized images for downstream classification tasks. This joint optimization method considers the complexity of missing modalities within a unified framework, offering a new approach for incomplete modality and opening new paths for research and applications in medical imaging and neuroscience.

Classification Learning Sub-network: The main difference between this network and conventional networks is that it incorporates high-dimensional transformed features, which contain shared information from cross-modal neuroimaging synthesis, in the decision-making process. This has not been considered in previous works. The advantage of this approach is that it can effectively extract shared information for decision-making. On this basis, the construction of each modality network no longer needs to consider the extraction of shared features, making the model construction easier and improving classification performance. The image synthesis sub-network and the classification learning sub-network are not independent; the transformation features from the MRI-to-PET cross-modal image generator and the classification learning sub-network are shared, as are the transformation features from the PET-to-MRI cross-modal image generator and the classification learning sub-network. These two transformation features are involved in both image synthesis and classification learning tasks. During image synthesis, the gradient feedback from the classification learning sub-network can guide the image generation of more disease-related missing modality images. During classification learning, the transformation features involved in cross-modal image generation can provide shared information from the modality transformation process to participate in decision-making, thereby enhancing the model's generalization ability.

By extracting disease-related features from both MRI and PET modalities, the network captures modality-specific information related to disease diagnosis and classification. These modality-specific features provide valuable information for disease diagnosis and contribute significantly to the overall performance of classification tasks. Furthermore, the transformation features obtained through the image synthesis process serve as an important bridge between the synthesis and classification tasks. These transformation features represent cross-modality shared information and guide the fusion process. By enhancing classification learning performance with these transformation features, the Classification Learning (CL) sub-network incorporates both shared and private information, thereby improving the accuracy of the model's classification task. Considering both private features and transformation features enables a more comprehensive and effective analysis of the complex relationships between different modalities, thus enhancing the diagnostic ability of the JISCL framework.

3.2.1. Problem formulation:

Let $\mathbf{I}_M$ denote the domain of MRI modality and $\mathbf{I}_P$ be the domain of PET images. We denote a set of subjects (with all MRI and PET scans, missing modality is replaced with 0) as $D = \left\{ \left(\mathbf{X}_M, \mathbf{X}_P \right) \mid \mathbf{X}_M \in \mathbf{I}_M, \mathbf{X}_P \in \mathbf{I}_P \right\}$. The generator G_P takes an MRI $\mathbf{X}_M$ as input and generates the synthesised PET $\mathbf{X}_{P_{syn}}$ and the converted feature $\mathbf{C}_{M2P}$.

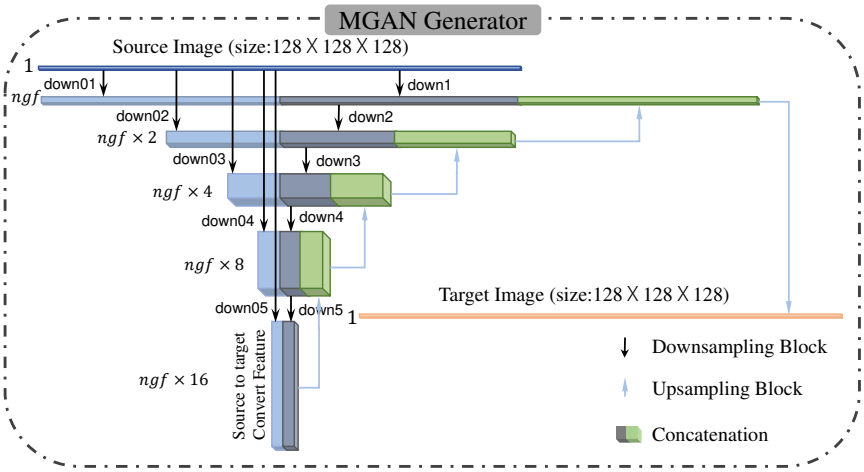


Figure 4: The network architecture of MGAN generator with converted features.

Similarly, the generator G_M takes a PET $\mathbf{X}_P$ as input and generates the synthesised MRI image $\mathbf{X}_{M_{syn}}$ and the converted feature $\mathbf{C}_{P2M}$. If $\mathbf{X}_M = 0$, we use $\mathbf{X}_{M_{syn}}$ as the input for generator G_P. If $\mathbf{X}_P = 0$, we use $\mathbf{X}_{P_{syn}}$ as the input for generator G_M.

By adopting this strategy, we can effectively utilize the synthesised modality information in the case of modality missing and incorporate it into the subsequent image generation process. This approach helps improve the performance of the generators and allows us to flexibly perform image synthesis in different modality missing scenarios. Both generators follow the MGAN generator structure shown in Figure 4. Specifically, each image generator consists of 10 downsampling blocks, as described in Table 2. Through the downsampling blocks, the input image size is reduced from $128 \times 128 \times 128$ to $4 \times 4 \times 4$ (converted feature), while the channel size is increased to $ngf \times 16 \times 2$. The upsampling blocks use `Upsample-conv3d-InstanceNorm3d-Relu`, and the final output layer uses `Tanh` instead of `Relu`. In the architecture, features are concatenated along the channel dimension and used for upsampling.

In our framework, feature fusion is performed using the extracted converted features $\mathbf{C}_{M2P}$ and $\mathbf{C}_{P2M}$. Instead of directly using synthesized MRI or PET images for fusion, we leverage the high-dimensional feature representations extracted from the generative networks. The converted features are concatenated along the channel dimension and input into a shared feature fusion module, which consists of a series of convolutional layers followed by batch normalization and ReLU activation. This ensures that the extracted features from different modalities are aligned in the feature space before classification.

To prevent the feature fusion process from degrading the quality of the synthesized images, we restrict the fusion operation to the classification branch and do not use the fused features to generate images. By designing the network in this way, we ensure that the converted features contribute only to the downstream classification task while maintaining the integrity of the synthesized images for cross-modality tasks.

For classifiers 1 and classifiers 2, we use the same classification structure. Specifically, each classifier consists of 4 downsampling blocks. In each downsampling step, a Conv3d is applied with a kernel size of 3, a stride of 1, and padding of 1. Subsequently, a $3 \times 3 \times 3$ AvgPool3d layer is used, and the output from the fourth layer is fed into a Linear layer for classification. For classifiers 3 and 4, we directly employ a Linear layer for classification.

In the JISCL framework, the classifier C plays a crucial role in determining the stage of Alzheimer's disease for patients. The true stage information of the patients is represented by the label y. During the classification task, the generated images are utilized for classification purposes only when either the MRI or PET modality is missing. In contrast, if both the MRI and PET modalities are complete and available, the classification task is performed using the real MRI and PET images instead of the generated ones. This design ensures that when complete modality data is accessible, the classification is carried out using real images, which enhances the accuracy and reliability of the classification task. By utilizing the generated images exclusively when modality data is missing, the framework leverages the benefits of cross-modal image synthesis while relying on real images for more comprehensive and accurate classification predictions.

Table 2
Network structure of the generator.

name	Layer	Kernel Size	Stride	Padding	Bias	Relu	attention	bias
down1-5	Conv3d	3	2	1	InstanceNorm3d	LeakyReLU	-	-
down01	Conv3d	3	2	1	InstanceNorm3d	-	SpatialAttention3d	InstanceNorm3d
down02	Conv3d	7	4	2	InstanceNorm3d	-	SpatialAttention3d	InstanceNorm3d
down03	Conv3d	25	7	1	InstanceNorm3d	-	SpatialAttention3d	InstanceNorm3d
down04	Conv3d	31	15	4	InstanceNorm3d	-	SpatialAttention3d	InstanceNorm3d
down05	Conv3d	55	27	4	InstanceNorm3d	-	SpatialAttention3d	InstanceNorm3d

3.2.2. Joint Image Synthesis and Classification Learning:

As shown in Figure 2, our IS model is based on the MGAN, which is designed to address the challenges of generating missing modalities and extracting converted features. The model consists of two generators (G_M, G_P) and two discriminators (D_M, D_P). The discriminator D_P takes a pair of images, including real image: $\left(\mathbf{X}_M, \mathbf{X}_P\right)$ and fake image (Use $[\mathbf{M}\cdot]$ instead of $\left[\mathbf{X}_M \text{ if } \mathbf{X}_M \neq 0 \text{ else } \mathbf{X}_{M_{syn}}\right]$): $\left([\mathbf{M}\cdot], G_P\left([\mathbf{M}\cdot]\right)\right)$, D_P aims to distinguish between real and synthesised images. The structure of the discriminator is designed based on pix2pix [16], using a `Conv3d-BatchNorm3d-LeakyReLU-Conv3d` structure and outputting a single channel for classification. The objective of the adversarial loss function is to minimize the difference between the real and synthesised images, encouraging the synthesised images to be more realistic. Additionally, we use a reconstruction loss function to measure the similarity between the synthesised images and the input images. By combining these loss functions, we are able to train the IS model to generate high-quality missing modality images and their corresponding converted features. We use the following adversarial loss to make the synthesised and real images distribution consistent.0..

$$L\left(D_P\right) = \sum_{\{X_M \in I_M, X_P \in I_P\}} \left(\alpha \cdot \beta \cdot \log\left(1 - D_P\left(X_M, X_P\right)\right) + \alpha \cdot \log D_P\left([\mathbf{M}\cdot], G_P\left([\mathbf{M}\cdot]\right)\right)\right), \tag{1}$$

$\alpha = 0$ when $\mathbf{X}_M = 0$, $\alpha = 1$ when $\mathbf{X}_M \neq 0$, $\beta = 0$ when $\mathbf{X}_P = 0$, and $\beta = 1$ when $\mathbf{X}_P \neq 0$. When $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$, both loss terms are calculated normally. When $\mathbf{X}_M = 0$ and $\mathbf{X}_P \neq 0$ ($\beta = 1$ and $\alpha = 0$), the loss is set to 0, and no gradient backward. When $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P = 0$ ($\beta = 0$ and $\alpha = 1$), only the loss of $G_P([\mathbf{M}\cdot])$ is computed. The D_M loss function is defined as equation (2).

$$L\left(D_M\right) = \sum_{\{\mathbf{X}_M \in \mathbf{I}_M, \mathbf{X}_P \in \mathbf{I}_P\}} \left(\alpha \cdot \beta \cdot \log\left(1 - D_M\left(\mathbf{X}_P, \mathbf{X}_M\right)\right) + \beta \cdot \log D_M\left([\mathbf{P}\cdot], G_M\left([\mathbf{P}\cdot]\right)\right)\right), \tag{2}$$

$[\mathbf{P}\cdot]$ is $\left[\mathbf{X}_P \text{ if } \mathbf{X}_P \neq 0 \text{ else } \mathbf{X}_{P_{syn}}\right]$. When $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$, both losses are calculated, when $\mathbf{X}_M = 0$ and $\mathbf{X}_P \neq 0$ ($\beta = 1$ and $\alpha = 0$), only the loss of $G_M([\mathbf{P}\cdot])$ is calculated and when $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P = 0$ ($\beta = 0$ and $\alpha = 1$), the loss is 0 and no gradient backward. The G_P loss function is defined as equation (3).

$$\begin{aligned} L\left(G_P\right) = \sum_{\{\mathbf{X}_M \in \mathbf{I}_M, \mathbf{X}_P \in \mathbf{I}_P\}} & \left(\alpha \cdot \beta \cdot \|G_P\left(\mathbf{X}_M\right) - \mathbf{X}_P\|_1 + \gamma \cdot \beta \|G_P\left(\mathbf{X}_{M_{syn}}\right) - \mathbf{X}_P\|_1 + \right. \\ & \left. \alpha \cdot \log\left(1 - D_P\left(G_P\left(\mathbf{X}_M\right)\right)\right) + \gamma \cdot \beta \cdot \log\left(1 - D_P\left(G_P\left(\mathbf{X}_{M_{syn}}\right)\right)\right)\right), \end{aligned} \tag{3}$$

Where $\|\cdot\|_1$ (denoting the l_1 norm) is the synthetic image loss, and the $\log(\cdot)$ is adversarial loss. When $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$, $\gamma = 0$, else, 1. Calculate the 1st and 3rd terms when $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$ ($\alpha = 1, \beta = 1, \gamma = 0$), the 2nd and 4th terms when $\mathbf{X}_M$ is missing ($\alpha = 0, \beta = 1, \gamma = 1$), and the 3rd term when $\mathbf{X}_P$ is missing ($\alpha = 1, \beta = 0, \gamma = 1$). The G_M loss function is defined as equation (4).

$$\begin{aligned} L\left(G_M\right) = \sum_{\{\mathbf{X}_M \in \mathbf{I}_M, \mathbf{X}_P \in \mathbf{I}_P\}} & \left(\alpha \cdot \beta \cdot \|G_M\left(\mathbf{X}_P\right) - \mathbf{X}_M\|_1 + \gamma \cdot \alpha \|G_M\left(\mathbf{X}_{P_{syn}}\right) - \mathbf{X}_M\|_1 + \right. \\ & \left. \beta \cdot \log\left(1 - D_M\left(G_M\left(\mathbf{X}_P\right)\right)\right) + \gamma \cdot \alpha \cdot \log\left(1 - D_M\left(G_M\left(\mathbf{X}_{P_{syn}}\right)\right)\right)\right), \end{aligned} \tag{4}$$

Calculate the 1st and 3rd terms when $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$ ($\alpha = 1, \beta = 1, \gamma = 0$), the 3rd terms when $\mathbf{X}_M$ is missing ($\alpha = 0, \beta = 1, \gamma = 1$), and the 2nd and 4th term when $\mathbf{X}_P$ is missing ($\alpha = 1, \beta = 0, \gamma = 1$). our modal classification model is formulated in equation (5):

$$y = C\left([\mathbf{M}\cdot], [\mathbf{P}\cdot], C_{M2P}, C_{P2M}\right), \tag{5}$$

Where y is the predicted class label (eg. CN\AD or sMCI\pMCI), and C is a classifier that can identify a patients stage. Use synthesised images for classification when MRI or PET is missing. Taking into consideration the unavoidable loss some features in synthesised images, we incorporated three sets of weights during the training of the classifier. For cases where both $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P \neq 0$, the weight for $[\mathbf{M}\cdot]$ is set to 0.34, and the weight for $[\mathbf{P}\cdot]$ is set to 0.36. When $\mathbf{X}_M = 0$ and $\mathbf{X}_P \neq 0$, the weight for $[\mathbf{M}\cdot]$ is adjusted to 0.2, while the weight for $[\mathbf{P}\cdot]$ is set to 0.5. Similarly, when $\mathbf{X}_M \neq 0$ and $\mathbf{X}_P = 0$, the weight for $[\mathbf{M}\cdot]$ is set to 0.5, and the weight for $[\mathbf{P}\cdot]$ is set to 0.2. The weights for C_{M2P} and C_{P2M} are set to 0.15 in all three cases.

In order to jointly optimize image synthesis and classification tasks, we integrated the image synthesis loss and classification loss. The total loss of JISCL is defined as $L(GC)$, see equation (6):

$$L(GC) = L\left(G_M\right) + L\left(G_P\right) + L(C). \tag{6}$$

During the training process, the optimization is performed in two steps.

First, the two discriminators (D_P and D_M) are optimized. The discriminators aim to distinguish between real and synthetic images. The optimization process involves updating the parameters of the discriminators to improve their ability to classify the images accurately. After optimizing the discriminators, the next step involves jointly optimizing the IS and CL networks while keeping the discriminators fixed. This means that during this phase, the parameters of the discriminators remain unchanged, and only the parameters of the IS and CL networks are updated.

By optimizing the IS and CL networks together, the model aims to improve the quality of the synthesised images and the accuracy of the classification task. The shared converted features play a crucial role in guiding the synthesis process and facilitating multi-modal fusion for classification. This joint optimization process helps to enhance the overall performance and accuracy of the JISCL framework.

Our approach incorporates several strategies that contribute to the diagnosis of AD. Firstly, the JISCL framework integrates the IS and CL sub-networks into a unified framework. This integration allows for joint optimization of cross-modality image synthesis and multi-modal classification, leveraging the benefits of both tasks to improve overall performance.

In the CL sub-network, we adopt a dynamic multi-modal image fusion strategy. When a modality is complete, we combine the complete modality with the converted features for classification prediction. This strategy maximizes the utilization of available modalities and enhances the accuracy of the classification task. On the other hand, when a modality is missing, we utilize the generated modality instead of the missing modality for image classification learning. This approach enables us to effectively handle the missing modality scenario and continue the learning process without discarding valuable information.

Additionally, the IS and CL sub-networks share the converted features. Instead of using separately extracted MRI/PET features, the CL network takes the converted features obtained from the IS network during image generation as inputs. By incorporating the converted features learned by the IS network, the CL network can better capture and retain the underlying relationships between multiple modalities from the same subject. This improves the accuracy of classification by leveraging the Synthesised information.

4. Experiments

4.1. Evaluation of generated images

We conducted a comparative analysis of different generative models, including conventional GAN with Resnet-6, CycGAN [29] using Resnet-6, pix2pix [16] using U-Net-128, and our proposed MGAN. To ensure a fair comparison, we set the number of channels in the first layer of the generator (ngf) to 32. We trained a total of 8 GAN models using the ADNI-2 dataset for training, which had a larger number of paired subjects, and evaluated their performance on the ADNI-1 dataset, which had fewer paired subjects. Three metrics, namely peak signal-to-noise ratio (PSNR) [30], structural similarity index measure (SSIM) [31], and normalized root-mean-square error (NRMSE) [32], were employed to assess the quality of the synthesized images.

Table 3
Results of Image Synthesis Achieved by four Different Methods for MRI and PET of Subjects in ADNI-1, With the Models Trained on ADNI-2.

Method	synthesis MRI			synthesis PET		
	PSNR	SSIM (%)	NRMSE	PSNR	SSIM (%)	NRMSE
GAN	27.85	**87.47**	1.24	25.48	91.47	0.46
CycleGAN	27.57	85.34	1.03	23.48	87.55	0.58
Pix2pix	27.24	86.12	1.25	25.69	91.72	0.46
MGAN(ours)	**28.41**	86.03	**0.98**	**25.82**	**91.82**	**0.45**

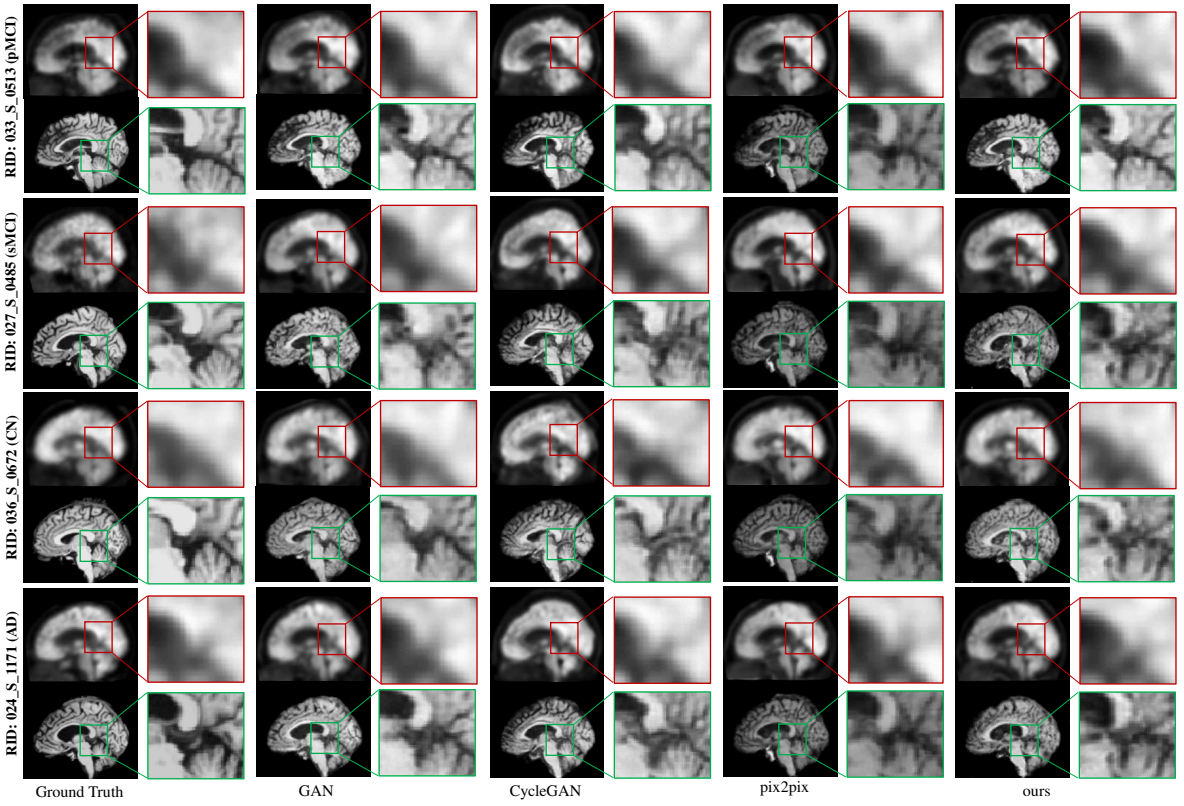


Figure 5: MRI and PET scans synthesis by four methods for four typical subjects in ADNI-1, along with corresponding ground-truth images. All four synthesis models are trained on ADNI-2.

Table 3 reveals interesting findings. Firstly, the MRI-to-PET synthesis task exhibits lower PSNR values but higher SSIM values compared to the PET-to-MRI task. This suggests that MRI excels in preserving structural information, while PET faces challenges due to misalignment with tissue structures. Among the four approaches, our method demonstrates superior performance in image quality metrics for synthesised PET, whereas the SSIM metric for synthesised MRI is comparatively lower compared to the GAN and pix2pix methods. This disparity can be attributed to the fact that our multi-scale generative adversarial network did not leverage guided converted features at this stage.

To further refine the effectiveness of our synthesis approach, we emphasize the importance of feature selection after feature fusion. While multi-modal feature fusion enhances the synthesis process by combining complementary information, not all features contribute equally to the final image quality. We adopt a feature selection strategy that prioritizes the most informative features, reducing redundancy and mitigating noise introduced during the fusion process. This targeted selection ensures that only the most relevant features guide the synthesis, improving both the realism and diagnostic utility of the generated images.

Figure 5 presents a visual comparison of real and synthesised images for AD (Roster ID: 1171), CN (Roster ID: 0672), sMCI (Roster ID: 0485), and pMCI (Roster ID: 0513) on ADNI-1. The enlarged red/green regions on the right side allow for better observation of image details. Our MGAN (5th column) demonstrates superior performance in generating synthesised images that closely resemble the real images (1st column), particularly in terms of ventricle size, compared to other GANs (2nd-4th columns). This improvement can be attributed to the multi-scale information extraction capability of our MGAN, which goes beyond voxel-level correspondence and encourages structural similarity.

Upon closer examination of the image details, we notice that the synthesised PET images retain certain characteristics from the original MRI images and exhibit differences from real PET images. This discrepancy arises from the fact that metabolic conditions detected by PET do not align precisely with tissue structures. Additionally, the synthesised MRI images appear slightly blurred and less structurally clear compared to the original images. This may be attributed to the inherent lack of well-defined boundaries in PET scans. These observations underscore the inherent challenges involved in cross-modality synthesis.

Our findings indicate that a well-defined feature selection mechanism within the fusion process is crucial for ensuring the robustness of cross-modality synthesis. By selectively extracting the most relevant features from both modalities, we can enhance the fidelity of generated images while preserving disease-specific biomarkers, leading to improved diagnostic accuracy.

4.2. Evaluation of incomplete modality disease diagnosis

4.2.1. Competing Methods:

To evaluate the performance of the JISCL model, we conducted experiments on two classification tasks: AD recognition and MCI conversion prediction. We compared the JISCL model with a traditional approach that involves generating missing images followed by multi-modal fusion. For image generation, we utilized the MGAN model, which achieved the best results in our experiments. For fusion, we employed three fusion methods: (1) decision-level fusion (DF) [33]. DF: Aggregates predictions from individual models at different stages or levels to effectively capture complementary information and improve classification performance. (2) feature-level fusion (FF) [34]. FF: Integrates information from different modalities by combining or merging extracted features to obtain a more comprehensive and accurate representation. and (3) low-rank multi-modal fusion (LWF) [35]. LWF: Exploits the low-rank structure of data matrices to capture shared information across modalities and extract a common low-dimensional representation for fused information.

To ensure fairness, we performed separate experiments on complete and incomplete modalities to assess the effectiveness of modality completion and joint optimization. We compared the performance of the JISCL model with the traditional approach using task-specific evaluation metrics.

4.2.2. Experimental Setup:

In terms of prediction experiments, we performed classification tasks for AD versus CN and pMCI versus sMCI to predict conversion in patients with MCI. To evaluate the performance of disease diagnosis, we used six metrics: (1) Area Under the Curve (AUC), (2) Accuracy (ACC), (3) Sensitivity (SEN) [36], (4) Specificity (SPE) [36], (5) F1-Score (F1S) [37], and (6) Matthews Correlation Coefficient (MCC) [38]. These metrics were employed to assess the performance of each method in the task of disease diagnosis.

4.2.3. Disease diagnostic results on ADNI-2:

The disease classification results obtained by different classification methods, trained on the ADNI-1 dataset and tested on the ADNI-2 dataset, are presented in Table 4. When trained on paired data, our JISCL achieves optimal performance in most cases. In the diagnosis of AD and CN, our JISCL achieves the best performance (except for AUC). In the diagnosis of pMCI and sMCI, it achieves the highest F1S (58.95%) and SPE (83.76%). This indicates that introducing converted features in JISCL improves the diagnostic performance of the model, Although in the presence of complete modalities. This indicates that the introduced converted features in JISCL also have the advantage of improving model performance in the presence of complete modalities.

After completing the missing modalities, there is a notable decrease in diagnostic performance for the two-stage methods (DF, FF, LWF). All evaluation metrics demonstrate a significant decline compared to the scenario without using synthesised neuroimaging data. In the AD/CN diagnostic task, the ACC shows a minimum decrease of 7.33% (FF) and a maximum decrease of 20.79% (LWF). For the MCI conversion diagnostic task, the ACC experiences a

Table 4
Disease classification results (%) by four different classification methods trained on ADNI-1 and tested on ADNI-2

Method	AD/CN Classification						pMCI/sMCI Classification					
	ACC	AUC	F1S	SPE	SEN	MCC	ACC	AUC	F1S	SPE	SEN	MCC
DF-C	90.29	94.41	86.25	**93.06**	85.29	78.75	**76.83**	75.82	58.70	81.48	63.53	42.96
FF-C	90.55	95.27	87.05	91.43	88.97	79.66	74.09	70.01	49.7	82.72	49.41	32.25
LWF-C	89.50	93.57	84.96	**93.06**	83.09	76.95	73.48	65.64	48.52	82.3	48.24	30.66
ours-C	**90.71**	93.92	**8**8.11	90.35	**91.30**	80.64	76.07	72.32	58.95	**83.76**	60.87	42.13
*mDSNet-C	90.64	**96.31**	**89.42**	91.39	89.70	**81.03**	76.23	**81.84**	**61.68**	76.95	**74.16**	**46.52**
DF	77.55	82.44	68.17	80.55	71.62	51.03	71.55	65.15	45.20	80.00	46.51	26.03
FF	83.22	85.13	75.50	86.35	77.03	62.77	73.31	60.35	40.52	85.88	36.05	23.97
LWF	68.71	67.12	56.05	73.38	59.46	32.01	67.16	60.71	44.55	72.16	52.33	22.44
ours	95.02	95.59	92.99	94.48	**96.05**	89.25	**80.12**	72.22	59.52	**90.32**	53.19	47.18
*mDSNet	93.05	**97.23**	92.02	94.74	90.91	85.88	79.71	**84.44**	**65.69**	81.25	**75.28**	**52.47**
ours-C_{M2P}	**95.48**	96.15	**93.42**	**96.55**	93.42	**89.97**	78.95	78.79	57.14	84.09	61.54	43.48
ours-C_{P2M}	93.21	96.93	89.66	95.27	89.04	84.61	79.53	75.53	63.92	84.68	65.96	49.70
ours-$(C_{M2P}+C_{P2M})$	93.67	95.45	91.14	93.1	94.74	86.37	77.78	75.67	63.46	80.65	70.21	48.16

-C: experiments conducted on fully aligned and paired MRI and PET scans. -C_{M2P}: The converted feature C_{M2P} is not utilized in the classification network. -C_{P2M}: The converted feature C_{P2M} is not utilized in the classification network. -$(C_{M2P}+C_{P2M})$: The converted feature C_{M2P} and C_{P2M} is not utilized in the classification network. *:The experimental results are cited from the paper [25], with a small difference in the number of datasets (in the single digits), and the incomplete modality was supplemented only with PET data.

decrease of at least 0.78% (FF) and up to 6.32% (LWF). These results indicate that treating cross-modality image generation and multi-modal fusion as separate tasks is inefficient and impairs diagnostic performance by hindering the learning of synthesised images aligned with downstream tasks. The decline in performance can be attributed to the loss of diagnostic features in the synthesised images and the lack of distinction between synthesised and real images in the two-stage approach, significantly affecting the model's learning process. In our model, there is only a slight decrease in the AUC and SEN metrics for the sMCI and pMCI classification tasks, while all other evaluation results show significant improvements. This indicates that our proposed method effectively addresses the problem of incomplete modalities, and the supplemented data is effectively used for diagnosis.

In the conducted ablation study, it was interesting to observe that the optimal results were not always achieved when both converted features (C_{M2P} and C_{P2M}) were used. In the diagnosis of AD and CN, utilizing C_{P2M} alone showed a slight advantage over using both converted features. The model achieved the highest performance with an ACC of 95.48%, F1S of 93.42%, SPE of 96.55%, and MCC of 89.97% when utilizing the converted feature C_{P2M} alone. One possible explanation is that the PET-to-MRI synthesis task can achieve satisfactory results even without considering the diagnostic outcomes. Moreover, the MRI-to-PET task consistently demonstrated higher PSNR values compared to the PET-to-MRI task 4.1. The AD/CN diagnosis task exhibited a higher discriminative capacity compared to the MCI conversion task, leading to more significant differences in the synthesised neuroimaging results between these two groups. In the diagnosis of sMCI and pMCI, using both converted features achieved the highest ACC and SPE, but with lower SEN values. This could be attributed to the difficulty in distinguishing MCI cases and the challenges of simultaneously learning both converted features. Overall, the JISCL model proves to be effective in addressing the issue of incomplete modalities while achieving good diagnostic performance, making it highly valuable for solving real-world medical problems.

4.2.4. Exploration of Converted Feature Selection:

In Table 4, we observed that excluding the converted feature C_{M2P} led to a slight improvement in AD/CN classification performance. This raises the question of whether a single-feature model might be more effective or if specific converted features should be tailored to different classification tasks.

Additionally, we compared our method with DSNet to assess its feature extraction capabilities. While DSNet achieves competitive performance in AD/CN classification, it underperforms in pMCI/sMCI classification. This suggests that although DSNet effectively captures spatial disease-relevant features, our method benefits from cross-modal feature conversion, making it more robust in handling incomplete modalities.

Table 5
Disease classification results (%) modality dropout is at 20% trained on ADNI-1 and tested on ADNI-2

Dropout Rate	Method	AD/CN Classification						pMCI/sMCI Classification					
		ACC	AUC	F1S	SPE	SEN	MCC	ACC	AUC	F1S	SPE	SEN	MCC
MRI(-20%)	ours-F	95.02	95.59	92.99	94.48	96.05	89.25	**80.12**	72.22	59.52	**90.32**	53.19	47.18
	ours	**92.31**	94.14	**88.74**	94.46	88.16	**82.90**	80.07	**75.72**	**62.07**	89.52	57.45	**49.53**
	ours-C_{M2P}	91.40	93.44	87.42	93.79	86.84	80.89	78.95	75.14	61.70	85.48	**61.70**	47.19
	ours-C_{P2M}	91.86	93.75	88.00	**94.48**	86.84	81.85	77.78	68.69	51.28	85.61	51.28	36.89
	ours-$(C_{M2P}+C_{P2M})$	91.40	**94.19**	87.90	91.72	**90.79**	81.34	77.78	73.66	52.50	84.85	53.85	38.03
PET(-20%)	ours	**92.31**	**94.55**	**88.99**	**93.79**	**89.47**	**83.01**	**78.36**	71.45	53.16	85.61	53.85	39.10
	ours-C_{M2P}	91.40	94.24	87.42	93.74	86.84	80.89	76.61	77.72	**56.52**	79.55	**66.67**	**41.93**
	ours-C_{P2M}	90.05	93.11	85.71	91.72	86.84	78.09	77.19	79.16	56.18	81.06	64.10	41.66
	ours-$(C_{M2P}+C_{P2M})$	90.95	93.13	86.67	93.69	85.53	79.84	**78.36**	**71.50**	51.95	**86.36**	51.28	37.99

ours-F: The classification results when the test set is complete. -C_{M2P}: The converted feature C_{M2P} is not utilized in the classification network. -C_{P2M}: The converted feature C_{P2M} is not utilized in the classification network. -$(C_{M2P}+C_{P2M})$: The converted feature C_{M2P} and C_{P2M} is not utilized in the classification network.

To further evaluate the impact of missing modalities, we conducted experiments with a 20% dropout rate for both MRI and PET modalities, as shown in Table 5. These results indicate that DSNet, relying solely on spatial features, lacks the adaptability of our method when handling missing modalities, whereas explicit feature conversion enhances classification robustness.

When the MRI dropout rate is set to 20%, the model using both converted features achieves the best performance in AD/CN classification, while single-feature models exhibit a decline in overall metrics. The advantage of using both converted features is even more pronounced in pMCI/sMCI classification, where models without converted features perform the worst.

Similarly, when the PET dropout rate is set to 20%, the results remain consistent with the MRI dropout scenario. The most significant drop in accuracy occurs when C_{P2M} is excluded or when no converted features are used, suggesting that C_{M2P} alone may not be sufficient when PET dropout is high. In the pMCI/sMCI classification task, overall performance is poorer when PET is missing compared to MRI. The accuracy remains comparable whether converted features are used or not, with a decline in performance when only a single converted feature is utilized. These findings highlight that in more challenging classification tasks, the absence of PET images poses a significant limitation in MCI classification.

4.2.5. Error Analysis

While our MGAN model demonstrates superior performance in cross-modality neuroimaging synthesis, certain limitations remain, particularly in cases involving missing data, anatomical variations, and fine-grained structural details.

Firstly, when conducting experiments on the original dataset, we observed that in the AD/CN classification task, the best performance was achieved when the converted feature C_{M2P} was not used. A closer analysis of the ADNI-2 dataset reveals that PET images for the CN category are missing in 45 cases, while 10 cases are missing in the AD category. Since the missing PET data is compensated by image generation, it inevitably introduces noise, making the converted feature less effective in capturing shared information. Additionally, due to the limited size of the training set, the model struggles with generalization, leading to performance degradation.

To further examine this phenomenon, we controlled the modality dropout rate in our experiments. The results indicate that, overall, using both converted features (C_{M2P} and C_{P2M}) provides the best performance. This suggests that while individual converted features might introduce inconsistencies in certain tasks, their combined use enhances robustness by leveraging complementary information.

Another key challenge is that PET images synthesized from MRI scans sometimes fail to accurately capture subtle metabolic variations crucial for early-stage disease detection. This is primarily due to the inherent differences between MRI, which provides structural information, and PET, which captures functional metabolic activity. Similarly, synthesized MRI images tend to appear slightly blurred compared to their real counterparts, likely due to the lower spatial resolution and less distinct boundaries of PET scans.

Additionally, we observed that in cases with significant anatomical abnormalities, such as enlarged ventricles or severe cortical atrophy in AD patients, the synthesized images tend to deviate more from the ground truth. This suggests that the generative model struggles to handle extreme anatomical variations, possibly due to the limited representation of such cases in the training dataset. The high inter-subject variability further exacerbates this issue, leading to inconsistencies in generated images, especially for individuals with atypical brain structures.

Another notable limitation is the misalignment of fine-grained details between synthesized and real images. Although our MGAN architecture encourages structural similarity, voxel-level inconsistencies persist. This issue could be addressed by refining the loss function to emphasize local texture preservation or integrating anatomical priors to improve synthesis fidelity.

Overall, our findings suggest that while MGAN effectively generates high-quality cross-modality neuroimaging data, its performance is influenced by dataset limitations, missing modality challenges, and extreme anatomical variations. Future work could focus on incorporating larger and more diverse datasets, leveraging hybrid loss functions that combine pixel-wise and feature-based constraints, and introducing anatomical consistency constraints to further improve synthesis robustness and clinical applicability.

5. Discussion and Conclusion

In this study, we propose an approach for diagnosing AD based on incomplete neuroimaging modalities. The inherent complexities and uncertainties associated with modality incompleteness, particularly in the early stages of AD, make this task exceedingly challenging. By jointly optimizing image synthesis and classification, our method effectively leverages the synergy between generated features and diagnostic predictions, leading to enhanced performance and deeper insights into AD pathology. This holistic approach fosters the generation of realistic and informative images while ensuring alignment with diagnostic objectives, ultimately advancing the field of neuroimaging-based AD research.

Through extensive experiments and evaluations, we demonstrated that the JISCL model outperforms traditional approaches in AD diagnosis. By combining image synthesis and AD classification, our model addresses the challenges posed by incomplete modalities and achieves promising results in this challenging task. We observed that the performance of the JISCL model was superior to that of two-stage methods, reinforcing the effectiveness of a joint learning strategy that optimizes both cross-modality image synthesis and disease classification in a unified framework. The model successfully learned to generate neuroimaging data that aligns well with downstream diagnostic tasks, resulting in improved diagnostic performance.

There are several considerations for future research in the field. Firstly, it would be valuable to explore more advanced generative networks that are specifically tailored to different modality conversion tasks. For instance, incorporating techniques like StyleGAN [39] or Diffusion GAN [40] could potentially enhance the quality and realism of cross-modality image synthesis. These advanced models have shown promising results in various image generation tasks and could be adapted to improve the synthesis of medical images.

Additionally, integrating genetic modalities into the diagnostic process is another avenue for future exploration. Genetic biomarkers, such as *APOE* $\varepsilon 4$, are closely linked to AD progression and could provide complementary diagnostic insights beyond neuroimaging. By incorporating genetic modalities into the analysis, we could potentially improve the accuracy and reliability of disease diagnosis and prognosis. One promising approach is to utilize multimodal fusion techniques, such as attention-based deep learning models, to learn cross-domain interactions between neuroimaging features and genetic data. This would allow for a more holistic understanding of disease progression and risk factors.

Lastly, it is crucial to conduct extensive validation and evaluation studies on large-scale and diverse datasets to validate the effectiveness and generalizability of the proposed methods. Robust and reliable evaluation metrics should be utilized to assess the performance of the models accurately. Future efforts should also focus on real-world clinical deployment, ensuring the model's robustness in handling heterogeneous patient data from different imaging centers.

In conclusion, our study demonstrates the effectiveness of the proposed joint image synthesis and classification framework for diagnosing Alzheimer's disease using converted features from incomplete neuroimaging modalities. By integrating image synthesis and classification into a unified model, we address the challenges posed by modality incompleteness and achieve improved diagnostic performance. The introduction of converted features enhances information preservation across modalities, facilitating more robust feature extraction for disease classification. Our findings underscore the significance of considering both image generation and diagnostic tasks holistically rather than treating them as independent processes. Future research directions include the refinement of feature fusion strategies,

the development of more expressive generative models, and the incorporation of additional modalities, such as genetic and clinical data, to further enhance the diagnostic capabilities of the proposed framework.

6. Acknowledgment

The author would like to thank the author of the comparison methods for the code and real world data sets provided. This work was supported in part by Yunnan provincial major science and technology special plan projects: digitization research and application demonstration of Yunnan characteristic industry, under Grant: 202002AD080001. Natural Science Foundation of China (NSFC) under Grant No. 61876166, No.61663046, No.62061050, the Postgraduate Practice Innovation Foundation of Yunnan University ZC-23235167, and Young Scientists Fund of the NSFC (No.62301452).

References

[1] Clive Ballard, Serge Gauthier, Anne Corbett, Carol Brayne, Dag Aarsland, and Emma Jones. Alzheimer's disease. *the Lancet*, 377(9770):1019–1031, 2011.

[2] Kaj Blennow, Mony J de Leon, and Henrik Zetterberg. Alzheimer's disease. *The Lancet*, 368(9533):387–403, 2006.

[3] Philip Scheltens, Kaj Blennow, Monique MB Breteler, Bart De Strooper, Giovanni B Frisoni, Stephen Salloway, and Wiesje Maria Van der Flier. Alzheimer's disease. *The Lancet*, 388(10043):505–517, 2016.

[4] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. Alzheimer's disease. *The Lancet*, 397(10284):1577–1590, 2021.

[5] Zhaodong Chen, Fengtao Nan, Yun Yang, Jiayu Wang, and Po Yang. Analysing and evaluating complementarity of multi-modal data fusion in ad diagnosis. In *2022 18th International Conference on Mobility, Sensing and Networking (MSN)*, pages 835–840, 2022.

[6] Fengtao Nan, Shunbao Li, Jiayu Wang, Yahui Tang, Jun Qi, Menghui Zhou, Zhong Zhao, Yun Yang, and Po Yang. A multi-classification accessment framework for reproducible evaluation of multimodal learning in alzheimer's disease. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[7] Fengtao Nan, Yahui Tang, Po Yang, Zhenli He, and Yun Yang. A novel sub-kmeans based on co-training approach by transforming single-view into multi-view. *Future Generation Computer Systems*, 125:831–843, 2021.

[8] Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage*, 61(3):622–632, 2012.

[9] Timothy C. Durazzo, Niklas Mattsson, and Michael W. Weiner. Smoking and increased alzheimer's disease risk: A review of potential mechanisms , for the alzheimer's disease neuroimaging initiative. *Alzheimer's Dementia*, 10(Supplement):S122–S145, 2014.

[10] Vince D. Calhoun and Jing Sui. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(3), 2016.

[11] Mingliang Wang, Daoqiang Zhang, Dinggang Shen, and Mingxia Liu. Multi-task exclusive relationship learning for alzheimers disease progression prediction with longitudinal data. *Medical image analysis*, 53:111–122, 2019.

[12] Benjamin M Marlin. Missing data problems in machine learning. 2008.

[13] Kim-Han Thung, Chong-Yaw Wee, Pew-Thian Yap, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage*, 91:386–400, 2014.

[14] Yawen Huang, Leandro Beltrachini, Ling Shao, and Alejandro F Frangi. Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging. In *Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 118–126. Springer, 2016.

[15] 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.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[17] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 455–463. Springer, 2018.

[18] Shengye Hu, Baiying Lei, Shuqiang Wang, Yong Wang, Zhiguang Feng, and Yanyan Shen. Bidirectional mapping generative adversarial networks for brain mr to pet synthesis. *IEEE Transactions on Medical Imaging*, 41(1):145–157, 2021.

[19] Xiaobin Hu, Ruolin Shen, Donghao Luo, Ying Tai, Chengjie Wang, and Bjoern H Menze. Autogan-synthesizer: Neural architecture search for cross-modality mri synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 397–409. Springer, 2022.

[20] Yan Wang, Luping Zhou, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, and Dinggang Shen. 3d auto-context-based locality adaptive multi-modality gans for pet synthesis. *IEEE transactions on medical imaging*, 38(6):1328–1339, 2018.

[21] Muhammad Sajjad, Farheen Ramzan, Muhammad Usman Ghani Khan, Amjad Rehman, Mahyar Kolivand, Suliman Mohamed Fati, and Saeed Ali Bahaj. Deep convolutional generative adversarial network for alzheimer's disease classification using positron emission tomography (pet) and synthetic data augmentation. *Microscopy Research and Technique*, 84(12):3023–3034, 2021.

[22] 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 *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 80–89. Springer, 2018.

[23] Chia-Hsiang Kao, Yong-Sheng Chen, Li-Fen Chen, and Wei-Chen Chiu. Demystifying t1-mri to fdgˆ 18 18-pet image translation via representational similarity. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 402–412. Springer, 2021.

[24] Yunbi Liu, Ling Yue, Shifu Xiao, Wei Yang, Dinggang Shen, and Mingxia Liu. Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages. *Medical image analysis*, 75:102266, 2022.

[25] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021.

[26] Kerstin Kläser, Thomas Varsavsky, Pawel Markiewicz, Tom Vercauteren, Alexander Hammers, David Atkinson, Kris Thielemans, Brian Hutton, Manuel Jorge Cardoso, and Sébastien Ourselin. Imitation learning for improved 3d pet/mr attenuation correction. *Medical image analysis*, 71:102079, 2021.

[27] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[28] Florian Kurth, Christian Gaser, and Eileen Luders. A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (spm). *Nature protocols*, 10(2):293–304, 2015.

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

[30] Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 37–38. IEEE, 2012.

[31] Henry Rusinek, Mony J De Leon, Ajax E George, Leonidas A Stylopoulos, Ramesh Chandra, Gwenn Smith, Thomas Rand, Manuel Mourino, and Henryk Kowalski. Alzheimer disease: measuring loss of cerebral gray matter with mr imaging. *Radiology*, 178(1):109–114, 1991.

[32] Casper O da Costa-Luis and Andrew J Reader. Micro-networks for robust mr-guided low count pet imaging. *IEEE transactions on radiation and plasma medical sciences*, 5(2):202–212, 2020.

[33] Keke Tang, Yuexin Ma, Dingruibo Miao, Peng Song, Zhaoquan Gu, Zhihong Tian, and Wenping Wang. Decision fusion networks for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[34] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 903–907. IEEE, 2018.

[35] Qingzheng Wang, Shuai Li, Hong Qin, and Aimin Hao. Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis. *Information fusion*, 26:103–121, 2015.

[36] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

[37] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[38] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):1–22, 2021.

[39] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[40] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. 2022.