Morphological Feature Visualization of Alzheimer's Disease via Multidirectional Perception GAN

Wen Yu, Baiying Lei, *Senior Member, IEEE*, Shuqiang Wang, *Member, IEEE*, Yong Liu, Zhiguang Feng, Yong Hu, *Senior Member, IEEE*, Yanyan Shen, *Member, IEEE*, and Michael K. Ng, *Senior Member, IEEE*

***Abstract*—The diagnosis of early stages of Alzheimer's disease (AD) is essential for timely treatment to slow further deterioration. Visualizing the morphological features for early stages of AD is of great clinical value. In this work, a novel multidirectional perception generative adversarial network (MP-GAN) is proposed to visualize the morphological features indicating the severity of AD for patients of different stages. Specifically, by introducing a novel multidirectional mapping mechanism into the model, the proposed MP-GAN can capture the salient global features efficiently. Thus, using the class discriminative map from the generator, the proposed model can clearly delineate the subtle lesions via MR image transformations between the source domain and the predefined target domain. Besides, by integrating the adversarial loss, classification loss, cycle consistency loss, and *L*1 penalty, a single generator in MP-GAN can learn the class discriminative maps for multiple classes. Extensive experimental results on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate that MP-GAN achieves superior performance compared with the existing methods. The lesions visualized by MP-GAN are also consistent with what clinicians observe.**

***Index Terms*—Alzheimer's disease (AD), generative adversarial networks (GANs), lesion visualization, magnetic resonance (MR) images.**

I. Introduction

ALZHEIMER'S disease (AD) is an irreversible and chronic neurodegenerative disease with progressive impairment of memory and other mental functions. It is estimated to be the fifth leading cause of death in elderly people [1]. AD is caused by abnormal cell death in the brain, long before amnestic symptoms are observable [2]. The resulting brain atrophy is visible in structural magnetic resonance (MR) images. To date, AD is incurable but preventable. It is crucial to diagnose the early stages of AD by MR images for timely treatment [3]–[5]. Significant memory concern (SMC) and mild cognitive impairment (MCI) are the transitional stages between normal controls (NCs) and AD [6]. SMC and MCI present mild symptoms, and the disease-related regions are very subtle in MR images. Currently, the clinical diagnosis procedure is time-consuming and requires extensive clinical training and experience for clinicians. Thus, developing automatic methods using deep learning to visualize the brain changes for the early stages of AD is highly desirable. It can assist clinicians in AD analysis and may provide meaningful information on the pathogenesis of cognitive decline. However, this is a challenging task due to several reasons, such as low-intensity contrast between the lesion and other neighboring regions, indistinct boundary of the lesion, and irregular lesion shape.

To visualize features of different Alzheimer's stages in MR images, there already exist several feature visualization methods based on classification. These methods can be classified into two categories as follows.

1) The region of interest (ROI)-based classification approaches [1], [7]–[9] and patch-based classification approaches [10]. The performance of these methods is limited since the brain ROIs or patches need to be selected based on anatomical brain atlases or biological prior knowledge beforehand. Multiple steps are required to exact features from ROIs or patches for classification and subsequent visualization. Therefore, they tend to be sensitive to parameters and time-consuming.
2) Three strategies to visualize features for a convolutional neural network (CNN) classifier: a) by editing an input image and observing its effect on the prediction results, the occluded regions that have a significant impact on prediction can be visualized; b) by analyzing the gradients of the prediction for an input image, a heatmap can be produced for visualization; and c) by analyzing the activations of the feature maps for the image, the regions that are responsible for making the specific prediction can be visualized.

Manuscript received December 25, 2020; revised July 27, 2021; accepted September 26, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62172403 and Grant 61872351; in part by the International Science and Technology Cooperation Projects of Guangdong under Grant 2019A050510030; in part by the Distinguished Young Scholars Fund of Guangdong under Grant 2021B1515020019; in part by the Excellent Young Scholars of Shenzhen under Grant RCYX20200714114641211; in part by the Shenzhen Key Basic Research Projects under Grant JCYJ20200109115641762 and Grant JCYJ20180507182506416; and in part by the Hong Kong Research Grant Council (HKRGC) under Grant GRF 12200317, Grant 12300218, Grant 12300519, and Grant 17201020. *(Wen Yu and Baiying Lei contributed equally to this work.) (Corresponding authors: Shuqiang Wang; Yanyan Shen.)*

Wen Yu, Shuqiang Wang, and Yanyan Shen are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518060, China (e-mail: sq.wang@siat.ac.cn; yy.shen@siat.ac.cn).

Baiying Lei is with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518055, China (e-mail: leiby@szu.edu.cn).

Yong Liu is with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China.

Zhiguang Feng is with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China.

Yong Hu is with the Department of Orthopedics and Traumatology, The University of Hong Kong, Hong Kong.

Michael K. Ng is with the Department of Mathematics, The University of Hong Kong, Hong Kong.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3118369.

Digital Object Identifier 10.1109/TNNLS.2021.3118369

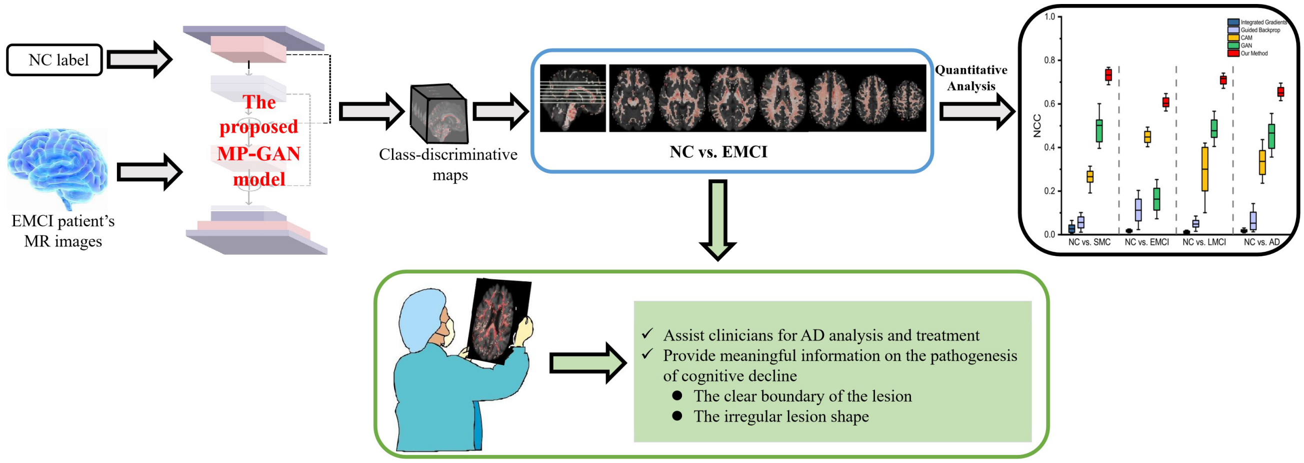

NC label
EMCI patient's MR images
The proposed MP-GAN model
Class-discriminative maps
NC vs. EMCI
Quantitative Analysis
NCC
1.0
0.8
0.6
0.4
0.2
0.0
Integrated Gradients
Guided Backprop
CAM
GAN
Our Method
NC vs. SMC
NC vs. EMCI
NC vs. LMCI
NC vs. AD
✓ Assist clinicians for AD analysis and treatment
✓ Provide meaningful information on the pathogenesis of cognitive decline
● The clear boundary of the lesion
● The irregular lesion shape

Fig. 1. Pipeline of the MP-GAN model. Assuming the MP-GAN is well-trained, given the real EMCI patient's MR images and NC label, the subtle morphological features between NC and EMCI can be visualized by the model to assist clinicians in AD analysis and treatment.

These classification-based feature visualization methods make their predictions based on local regions most relevant to particular prediction rather than the whole image, and it may ignore features with low discriminative power if stronger features for the prediction are available. As a result, if there is evidence for a category at multiple locations in the image (such as multiple AD lesions in MR images), some lesions with low discriminative power may be ignored. Moreover, visual features strongly depend on the performance of the classifier, and a large number of labeled samples are required to train a robust model.

To alleviate these issues, a novel multidirectional perception generative adversarial network (MP-GAN) is proposed to visualize morphological features for whole-brain MR images as shown in Fig. 1. GAN [11], [12] has attracted lots of attention as it is capable of generating realistic data without explicitly modeling the probability density function. Specifically, the generator of MP-GAN takes both MR images and its target domain as input. Then it flexibly learns a class discriminative map for the target domain. By adding the class discriminative map and the input MR image of the source domain, a synthetic MR image of the target domain can be produced. Thus, the learned class discriminative map can capture all the brain changes by transforming the MR image between the source domain and the target domain. By visualizing class discriminative maps, the subtle and complex lesions that may not be found within one region can be identified. Besides, by designing the hybrid loss, a single generator in MP-GAN can learn the class discriminative maps for multiple classes. In this manner, the common features unrelated to the specific domain can be reused during training, and therefore the visualization performance is further improved. With this global lesion visualization, clinicians can better exclude undesirable biases and potentially even identify previously unknown characteristics of AD. To the best of our knowledge, the proposed MP-GAN is the first work to visualize the morphological features for different Alzheimer's stages by a single generator. The contributions of this article are summarized as follows.

1) A novel MP-GAN with a multidirectional mapping mechanism is proposed to capture the salient global features efficiently. Using the class discriminative map from the generator, the proposed model can clearly delineate the subtle lesions via MR image transformations between the source domain and the target domain.
2) By integrating the adversarial loss, classification loss, cycle consistency loss, and $L1$ penalty, a single generator in MP-GAN can learn the class discriminative maps for multiple classes. The morphological features indicating different Alzheimer's stages can be visualized by a single MP-GAN model.

The rest of this article is organized as follows. The related work is reviewed in Section II. The proposed MP-GAN is described in detail in Section III. In Section IV, MP-GAN is tested and compared with the existing feature visualization methods to demonstrate its advantage. Finally, concluding remarks and future work are discussed in Sections V and VI.

II. Related Work

A. Generative Adversarial Networks

GAN has attracted lots of attention as it is capable of generating realistic data without explicitly modeling the probability density function. It has shown remarkable results in various computer vision tasks such as image generation [11], image-to-image translation [13], image super-resolution [12], and semisupervised learning [14], [15]. A typical GAN model consists of two modules: a discriminator and a generator. The discriminator learns to distinguish between real and fake samples, while the generator learns to generate fake samples that are indistinguishable from real samples. Training the original GAN, however, suffers from several problems such as low quality of generated images, convergence problems, and mode collapse. To address these deficiencies, variants of the GAN were introduced [16], [17]. The most representative work is Wasserstein GAN (WGAN) [16]. It leverages the Wasserstein distance to measure the distance between two data distribution that has better theoretical properties than the original Kullback–Leibler (KL) divergence.

B. Feature Visualization Methods

The current feature visualization methods for AD generally fall into two categories: 1) the ROI-based classification approaches and 2) the CNN-based classification approaches.

For the first category, the brain ROIs or patches were selected based on anatomical brain atlases or biological prior knowledge beforehand, and then multiple steps were required

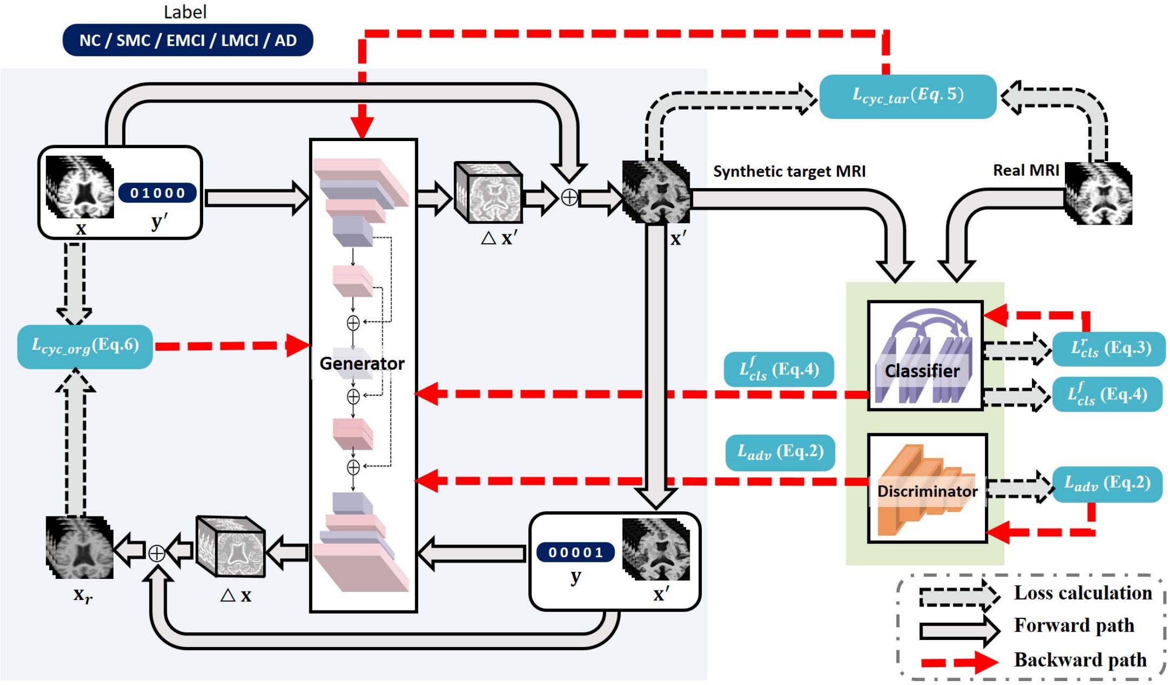


Fig. 2. Flowchart of MP-GAN. It consists of three components: a generator, a classifier, and a discriminator. The generator maps the input source MR image to the synthetic target MR images by the class discriminative map. The synthetic source MR images are reconstructed from the synthetic target MR images by the generator in the same manner. Using classification loss, adversarial loss, and cycle consistency loss, the generator learns to generate synthetic target MR images and the reconstructed source MR images that are indistinguishable from real MR images.

to extract features from ROIs or patches for classification. According to classification performance, the most frequently selected ROIs or patches would be visualized [7], [8]. For instance, Lian *et al.* [10] proposed a hierarchical fully convolutional network (H-FCN) to automatically identify discriminative local patches and regions in MR images for AD analysis. The hierarchical discriminative locations of brain atrophy at both the patch level and region level were visualized.

For the second category, there were three strategies to visualize features for CNN.

1) By editing an input image and observing its effect on prediction results, the occluded regions that had a significant impact on prediction can be visualized [18]. For instance, Zeiler and Fergus [19] proposed an occlusion-based method to visualize the activity within CNN. Different portions of the input image were occluded with a gray square, and the output of the classifier was observed. The occluded regions that cause the probability of the correct class drop significantly would be visualized. Korolev *et al.* [20] used 3-D-ResNet for AD classification, and the important regions of the MR image most affected by AD were visualized by the occlusion-based method [19].
2) By analyzing the gradients of the prediction for an input image, the heatmap can be produced for visualization [21]–[26]. For example, Springenberg *et al.* [27] proposed a new variant of the "deconvolution approach" guided backpropagation for visualizing features learned by CNNs. Guided backpropagation can be applied to a broader range of network structures. Sundararajan *et al.* [28] proposed integrated gradients using an axiomatic framework for feature visualization.
3) By analyzing the activations of feature maps for the image, the regions that were responsible for making specific prediction can be visualized. For instance, Zhou *et al.* [29] proposed class activation mapping (CAM) to visualize the discriminative object parts detected by CNN in a single forward pass. Khan *et al.* [30] used VGG with transfer learning for AD analysis. CAM was used to visualize the discriminative regions in the MR image for model interpretation. Lian *et al.* [31] proposed a multitask weakly supervised attention network (MWAN) by leveraging a fully trainable dementia attention block for regression. The attention maps were visualized by CAM for AD subjects. Sarraf and Tofighi [32] used LeNet-5 to classify structural MR images for AD versus NC. The filters and features were visualized for interpretation.

III. Proposed MP-GAN

A. Overview

This article proposes a novel mapping mechanism by which MR images can be mapped between each pair of source class and target class in a multidirectional manner. Take the source class NC as an example, the generator can map the MR images between NC and SMC; meanwhile, it can also map MR images between NC and the other class such as EMCI, LMCI, and AD simultaneously. The flowchart of MP-GAN is shown in Fig. 2. After data preprocessing (see Section IV-A), the normalized T1-MR images of all the classes are fed into MP-GAN. The proposed model learns the class discriminative maps between all class pairs for visualizing morphological features. More specifically, the generator aims to capture salient global features in class discriminative maps. Then the class discriminative maps are used to transform MR images between the source domain and the target domain. To control semantic information, an auxiliary classifier is introduced based on the generator and the discriminator to form the MP-GAN architecture. While the generator produces class discriminative maps distinguishing between the source domain and the target domain, the classifier predicts the domain indicating Alzheimer's stage, and the discriminator identifies whether the transformed MR images are real or fake. In this manner, the class discriminative maps learned by MP-GAN can highlight exactly which regions of the MR image are significant for discrimination between the source domain and the target domain at the voxel level. The subtle and complex lesions that may not be found within one region can be identified. Furthermore, since the input MR images are high-order with complicated brain structure, MP-GAN is further designed with the following two improvements: 1) 3-D residual blocks are exploited in the conditional generator so that features from low level can be reused, and the vanishing gradient problem can be prevented and 2) 3-D-DenseNet is used in the classifier to capture more discriminative features.

B. Architecture

The proposed MP-GAN is designed to visualize morphological features for multiple classes. To achieve this, the generator G is designed to produce class discriminative map Δx which can transform an input MR image x to an output MR image x' conditioned on the target class y', $[G(x, y')+x] \rightarrow x'$. During training, the target class y' is randomly selected so that G learns to produce class discriminative maps for all class pairs. By doing so, the target class y' can be predefined, and global features that distinguish between the source domain y and the desired target domain y' can be visualized at the testing stage.

As illustrated in Fig. 2, the input MR image x is labeled and y represents the corresponding class. The conditional generator aims to capture all the salient global features in class discriminative maps Δx. Then Δx is used to transform the input MR image from the source domain y to the target domain y' in a bidirectional manner. The classifier predicts label y_c given real MR image x by the conditional distribution $p_c(y|x)$, and the discriminator is trained to identify whether the MR image is real or fake. Formally, given an MR image x of source class y and a conditional variable y', the generator can produce a synthetic MR image x' of target class y' by adding the generated class discriminative map Δx and input MR image x

$$x' = \Delta x + x = G\big(x, y'\big) + x \tag{1}$$

which is indistinguishable from the real MR image of the target domain y'. Thereby, class discriminative map Δx contains all the salient global features that distinguish between two domains y and y'. The change in salient voxels between the source domain y and the target domain y' on the MR image can be visualized by the class discriminative map.

1) Adversarial Loss: To make the synthetic target MR images indistinguishable from real MR images, an adversarial loss is defined as

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_x\big[\log D(x)\big] + \mathbb{E}_{x,y'}\big[\log\big(1-D(G\big(x, y'\big)+x)\big)\big] \tag{2}$$

where generator G generates an MR image $[G(x, y') + x]$ conditioned on both the input MR image x and the target class y', while discriminator D attempts to distinguish between real and fake MR images. G tries to minimize this adversarial loss, while D tries to maximize it. More specifically, when the discriminator successfully identifies real and fake MR images, it is rewarded and no change is needed to update the parameters of the discriminator, whereas the generator is penalized with large updates to parameters. Alternately, when the generator fools the discriminator, it is rewarded, and no change is needed to update the parameters of the generator, but the discriminator is penalized and its component parameters are updated.

2) Classification Loss: Given an input MR image x and a target class y', the goal of MP-GAN is to produce a class discriminative map that can transform x into an output MR image x'. x' aims to be classified as the target class y'. To achieve this condition, an independent classifier is introduced and the classification loss is imposed when optimizing generator G. Specifically, the loss function is decomposed into two terms: a classification loss of real images to optimize classifier C and a classification loss of fake images to optimize generator G. In detail, the former is defined as

$$\mathcal{L}^r_{\text{cls}} = \mathbb{E}_{(x,y)\sim p_{\text{real}}(x,y)}\big[-\log p_c(y|x)\big]. \tag{3}$$

By minimizing this classification loss, classifier C learns to classify a real MR image x to its corresponding class y. On the other hand, the loss function for the classification of fake images is defined as

$$\mathcal{L}^f_{\text{cls}} = \mathbb{E}_{(x',y')\sim p_g(x,y)}\big[-\log p_c\big(y'|x'\big)\big]. \tag{4}$$

Generator G tries to minimize the loss $\mathcal{L}^f_{\text{cls}}$ to produce class discriminative maps for generating MR images x' that can be classified as the target class y'.

3) Cycle Consistency Loss: By minimizing the adversarial and classification losses, generator G is trained to generate MR images that are realistic and classified as target class. However, minimizing the losses [see (2) and (4)] does not guarantee that the final transformed images preserve the content of input MR images while changing only the disease-related regions of the input. To alleviate this problem, a forward cycle consistency loss and backward cycle consistency loss [13], [33] are applied to the generator. They are defined as

$$\mathcal{L}_{\text{cyc\_tar}} = \mathbb{E}_{x,y',y}\big[\big\|x'_{\text{real}} - \big(G\big(x, y'\big) + x\big)\big\|_1\big] \tag{5}$$

$$\begin{aligned}\mathcal{L}_{\text{cyc\_org}} &= \mathbb{E}_{x,y',y}\big[\big\|x_{\text{real}} - \big(G\big(x', y\big) + x'\big)\big\|_1\big] \\ &= \mathbb{E}_{x,y',y}\big[\big\|x_{\text{real}}-\big(G\big((G(x, y')+x), y\big)+x'\big)\big\|_1\big]\end{aligned} \tag{6}$$

where generator G takes in the transformed MR image x' and the source class y as input and tries to reconstruct the MR image $X_r = G\big(x', y\big) + x'$ of the source domain y. The $L1$ norm is adopted as the reconstruction loss. Note that a single generator is reused twice. The generator is first used to transform MR images from the source domain y to MR images of the target domain y'. Then it is used to reconstruct the MR image of the source domain y from the synthetic MR images of the target domain y'. For the first utilization, forward cycle consistency loss $\mathcal{L}_{\text{cyc\_tar}}$ is adopted. For the second one, backward cycle consistency loss $\mathcal{L}_{\text{cyc\_org}}$ is adopted.

4) L*1 Penalty:* The smallest class discriminative map Δx that leads to a real MR image of the target domain y' is encouraged. Thus, $L1$ penalty is defined as

$$\mathcal{L}_1(\Delta x) = \|\Delta x\|_1 \tag{7}$$

where $\|\cdot\|_1$ is the $L1$ norm.

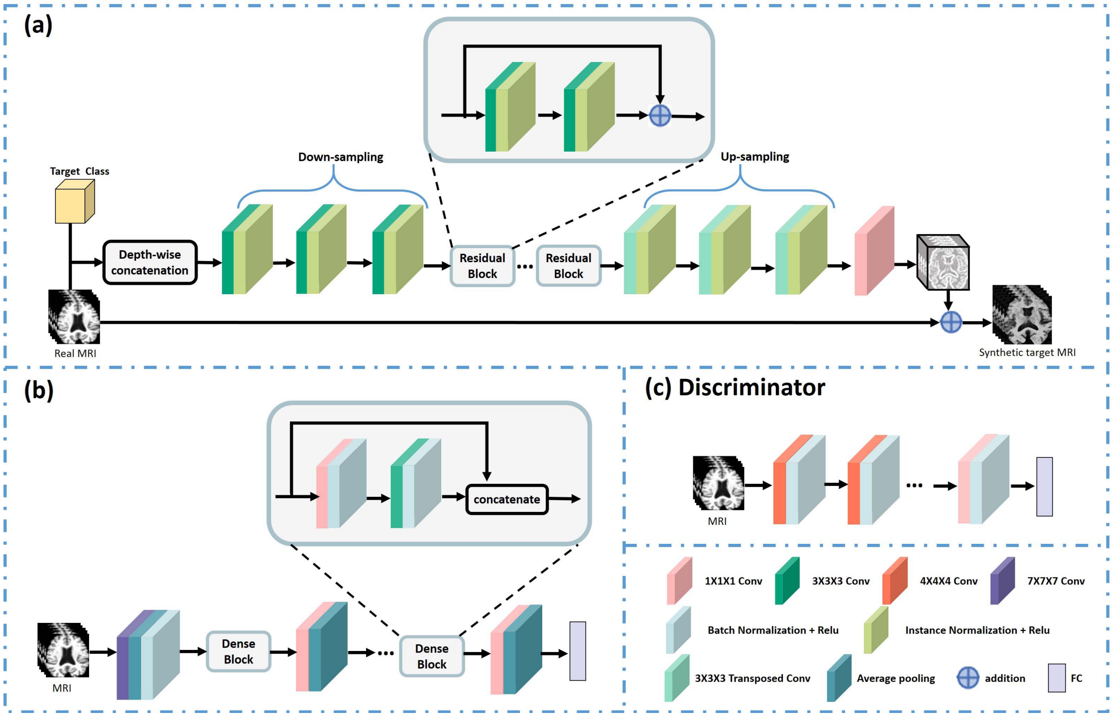


Fig. 3. Network architecture of the proposed MP-GAN. It consists of three components. (a) generator, (b) classifier, and (c) discriminator. The generator consists of two convolutional layers for downsampling, three 3-D-residual blocks, and two transposed convolutional layers for upsampling. 3-D-DenseNet is used in the classifier. A standard CNN architecture with seven convolutional layers with $4 \times 4 \times 4$ and $1 \times 1 \times 1$ convolutional filters is adopted in the discriminator.

5) Total Loss: The total loss functions to optimize D, C, and G are defined, respectively, as

$$\mathcal{L}_D = -\mathcal{L}_{\text{adv}} \tag{8}$$

$$\mathcal{L}_C = \mathcal{L}_{\text{cls}}^r \tag{9}$$

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}^f + \lambda_1\mathcal{L}_1(\Delta x) \\ + \lambda_{\text{cyc\_org}}\mathcal{L}_{\text{cyc\_org}} + \lambda_{\text{cyc\_tar}}\mathcal{L}_{\text{cyc\_tar}}\end{aligned} \tag{10}$$

where λ_{cls}, λ_1, $\lambda_{\text{cyc\_org}}$, and $\lambda_{\text{cyc\_tar}}$ are the hyperparameters that control the relative importance of classification loss, $L1$ penalty, and cycle consistency loss, respectively, compared with the adversarial loss. λ_{cls} is set as 0.1, λ_1 is set as 10, $\lambda_{\text{cyc\_org}}$ is set as 10, and $\lambda_{\text{cyc\_tar}}$ is set as 1 empirically throughout this article. Note that each loss term is indispensable for the proposed MP-GAN. Without any loss term of the hybrid loss, the training of MP-GAN will become extremely unstable and the learned class discriminative maps cannot capture the salient features for each pair of classes.

During the training process, the following domain settings are defined to train so that all the features between any two domains y and y' can be visualized for AD analysis.

(1) $y = \{\text{NC}\}, \quad y' = \{\text{SMC,EMCI,LMCI,AD}\}$.
(2) $y = \{\text{SMC}\}, \quad y' = \{\text{NC,EMCI,LMCI,AD}\}$.
(3) $y = \{\text{EMCI}\}, \quad y' = \{\text{SMC,NC,LMCI,AD}\}$.
(4) $y = \{\text{LMCI}\}, \quad y' = \{\text{SMC,EMCI,NC,AD}\}$.
(5) $y = \{\text{AD}\}, \quad y' = \{\text{SMC,EMCI,LMCI,NC}\}$.

From the algorithm perspective, take the first case (1) as an example, when source domain y is set as NC and the target domain y' will be set as one of {SMC, EMCI, LMCI, and AD}. Note that all the categories in {SMC, EMCI, LMCI, and AD} set will be trained at least once. In this way, all the above five conditions will be trained for MP-GAN, and thus a single generator in MP-GAN can learn the class discriminative maps for each pair of classes, and the salient global features can be captured for multiple classes. At the testing stage, y' is predefined according to the requirement of user. In this article, at the testing stage, the morphological features of NC versus all Alzheimer's stages including SMC, EMCI, and LMCI are visualized. MCI is characterized by a slight decline in cognitive abilities. Note that patients with MCI are at increased risk of developing AD, but do not always do. Thus, MCI is significant for morphological feature visualization and further AD analysis.

The network structure of the generator, classifier, and discriminator is shown, respectively, in Fig. 3. The network used in the generator is ResNet. 3-D-ResNet is expanded by adding a spatial dimension to all the convolutional and pooling layers in ResNet for the MR image. Using the shortcut connection, ResNet explicitly reformulates the layers as learning residual functions regarding the input layer, and it transfers feature representations from low layers to high layers. More specifically, assume that the target class y' is a discrete label, and it is encoded as a one-hot tensor. The target label y' is concatenated to the input MRI tensor in a depth-wise manner. Then they are operated by two convolutional layers with a stride size of 2 for downsampling, three 3-D-residual blocks [34], and two transposed convolutional layers with the stride size of 2 for upsampling. In this manner, the target label y' is operated with the input MRI tensor of the source label y in each hidden layer of the generator, and the class discriminative map Δx between y and y' will be generated.

TABLE I
DEMOGRAPHIC CHARACTERISTICS OF THE SUBJECTS IN ADNI DATASET

Magnet strength	3T								1.5T			
Source	ADNI-1		ADNI-GO	ADNI-2					ADNI-1		ADNI-GO	ADNI-2
Subject	NC	AD	EMCI	NC	SMC	EMCI	LMCI	AD	NC	AD	NC	NC
Number	42	29	142	190	121	309	177	159	171	175	16	81
Gender(F/M)	27 / 15	20 / 9	67 / 75	95 / 95	64 / 47	139 / 169	80 / 97	68 / 91	89 / 82	85 / 90	8 / 8	45 / 36
Age	76.1±5.1	75.6±7.9	71.7±7.7	74.9±6.8	72.9±5.6	72.3±7.3	72.9±7.6	75.4±7.9	77.7±5.4	76.6±7.5	80.2±4.8	82.5±4.5
Education	16±2.8	14.7±2.9	15.8±2.7	16.4±2.7	16.8±2.5	16±2.7	16.5±2.6	15.8±2.7	16.0±2.9	14.6±3.2	15.5±2.5	15.9±2.9
MMSE	29.3±1.0	20.03±4.8	28.2±1.8	28.7±1.5	28.6±1.7	28.0±2.1	26±3.5	20.8±4.4	29.1±1.2	21.5±4.4	29.4±1.0	28.5±2.6
CDR	0±0.14	1.07±0.4	0.45±0.19	0.07±0.19	0.13±0.23	0.46±0.22	0.58±0.37	0.99±0.46	0±0.19	0.93±0.49	0.07±0.18	0.2±0.35
Samples	149	73	471	723	288	1111	616	501	587	520	72	205

Finally, a synthetic MR image x' of target class y' is produced by adding the generated class discriminative map Δx and input MRI tensor x. Instance normalization [35] is used in all the layers except the last output layer for the generator. $3 \times 3 \times 3$ and $1 \times 1 \times 1$ convolutional filters are used in the generator. The network used in the classifier is DenseNet [36]. 3-D-DenseNet is expanded by adding a spatial dimension to all the convolutional and pooling layers in DenseNet for the MR image. Feature maps learned by all the preceding layers are concatenating along the last dimension for subsequent layers. Through such dense connectivity, feature maps are reused and the vanishing gradient problem is alleviated. Meanwhile, 3-D-DenseNet can extract discriminative features related to Alzheimer's stage from the whole MR images efficiently. The details of 3-D-denseNet can be found in [36] and [37]. In this article, the depth is set to 30, the growth rate is set to 12, the number of Dense-BC block is set to 3, and the reduction is set to 0.5. A standard CNN architecture with seven convolutional layers with $4\times4\times4$ and $1\times1\times1$ convolutional filters is adopted in the discriminator. Each convolutional layer is followed by batch normalization [38] and rectified linear unit (ReLU).

IV. EXPERIMENTS AND RESULTS

A. Dataset and Preprocessing

There are five stages associated with AD progression: NC, SMC, early MCI (EMCI), late MCI (LMCI), and AD. T1-weighted MR images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public dataset are used for evaluation purpose. In all, 5316 MR images in ADNI-1, ADNI-go, and ADNI-2 are used. It includes 1736 NC subjects, 288 SMC subjects, 1582 EMCI subjects, 616 LMCI subjects, and 1094 AD subjects. Both 1.5- and 3-T field strength MR images are used. Table I lists the demographic characteristics of the subjects.

ADNI-go and ADNI-2 added 129 and 782 participants, respectively, to the 819 recruited by ADNI-1.[1] In ADNI-1, NC and MCI participants continue to be followed by ADNI-go and ADNI-2. Different from the ADNI-1 dataset, MCI is divided into two subtypes, including EMCI, and LMCI in the ADNI-2 dataset. ADNI-go also added a new cohort of people with EMCI, and ADNI-2 added a cohort who were clinically evaluated as cognitively normal but had SMC. Note that SMC is the transitional stage between NC and MCI. The diagnostic criteria are described in the ADNI procedures' manual.[2]

All the MR images are in the neuroimaging informatics technology initiative (NIfTI) format. They are processed using standard operations in the FMRIB Software Library (FSL)[3] toolbox [39]–[41] for registering the MR images to the Montreal Neurological Institute (MNI) space. The preprocessing pipelines contain three parts: 1) removal of redundant tissues; 2) brain area extraction by BET; and 3) linear registration by FLIRT [42], [43]. Finally, the T1-MR image is normalized into the range [-1,1] and is fed into the MP-GAN model as a tensor directly without compressing or downsizing.

B. Experiment Settings

The proposed MP-GAN is trained on the ADNI dataset from scratch in an end-to-end manner. All the methods are implemented in TensorFlow.[4] All the experiments are conducted on four NVIDIA GeForce GTX 2080 Ti GPUs. "Adam" is used as the optimizer for stochastic gradient descent (SGD). The batch size is set to 8 empirically as each MR image is a high-order tensor of $109 \times 91 \times 91$. Since the batch size is relatively small, the gradients will become unstable, and thus there is a need to reduce the learning rate to stabilize training. According to the experimental results, the learning rate of both the generator and the classifier is set to 0.001, and the learning rate of the discriminator is set much smaller as 10^{-4}. For evaluation, 80% of the MR images are allocated for training. The remaining 20% of the MR images are equally partitioned and used as validation and test datasets, respectively. For avoiding bias, the training set, validation set, and test set do not have the MR images from the same subject simultaneously. A single MP-GAN model is trained on a training dataset of all the categories, and then the morphological features between the source domain and the predefined target domain are visualized on the test set. The validation dataset is used to tune hyperparameters to obtain the best model out of several epochs during the training process.

C. Qualitative Analysis

In this section, comprehensive experiments are conducted to show the effectiveness of MP-GAN. First, the proposed model is compared with four methods: 1) guided backpropagation [27]; 2) integrated gradients [28]; 3) CAM [29]; and 4) GAN [44]. For guided backpropagation, integrated gradients, and CAM, a conventional CNN architecture is used for these networks. More specifically, the CNN architecture consists of ten convolutional layers followed by batch normalization and max-pooling layers. After the last convolutional layer, an average pooling layer is used instead of the fully connected layer. Besides, for the CAM method, the last layer is designed as described in [29] and the last two max-pooling layers are omitted. This allows for more accurate heatmaps due to higher resolution of the last feature maps. The proposed method is also compared with the conventional GAN [44] to demonstrate our advantages. For a fair comparison, the network structure of the generator and the discriminator in

[1] http://adni.loni.usc.edu/about/

[2] http://www.adni-info.org

[3] www.fmrib.ox.ac.U.K./fsl

[4] http://www.tensorflow.org/

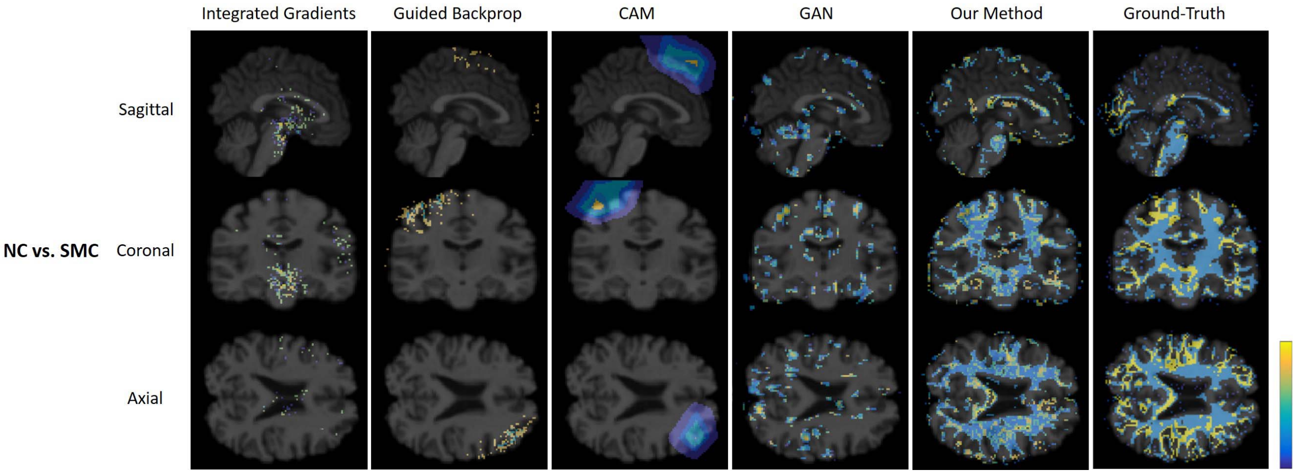


Fig. 4. Heatmaps predicted by integrated gradients, guided backprop, CAM, GAN, and our method are shown in sagittal, coronal, and axial views for NC versus SMC, respectively.

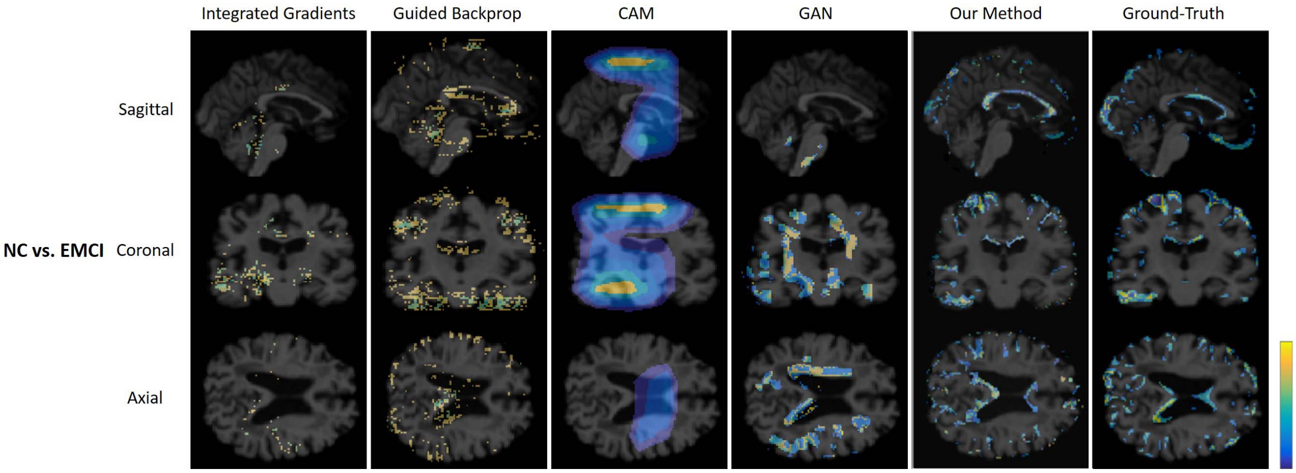


Fig. 5. Heatmaps predicted by integrated gradients, guided backprop, CAM, GAN, and our method are shown in sagittal, coronal, and axial views for NC versus EMCI, respectively.

GAN is the same as the proposed MP-GAN, and the loss function of GAN is the conventional adversarial loss. The GAN is trained to visualize important regions in MR images between two predefined classes. Furthermore, the following four evaluation groups are set up when compared with the four existing methods: 1) NC versus SMC; 2) NC versus EMCI; 3) NC versus LMCI; and 4) NC versus AD. The main reason for this setup is that more meaningful pathological features can be found by comparing with healthy people. It is worth noting that the MR images of all the five classes are trained using only one MP-GAN model, and the class discriminative map for each evaluation group is visualized at the test stage. But for the four compared methods, one independent binary model is trained for each evaluation group, respectively.

To visually show the quality of heatmaps produced by the proposed model and the four existing methods, one MR image is taken from each evaluation group for qualitative analysis. From Figs. 4–7, the heatmaps from the sagittal, coronal, and axial views are illustrated for each evaluation group, respectively. Figs. 4–7 are shown by progression from SMC to AD in order. From Figs. 4–7, it can be seen that the proposed MP-GAN can visualize subtle lesions with contour edge at a finer scale (i.e., voxel level). More detailed discriminative regions can be depicted, such as the hippocampus, and the corners and boundaries of the ventricle. The highlighted subtle lesions predicted by MP-GAN are relatively more precise than those generated by the other four methods. For example, from Fig. 5, it can be observed that the lesions that have much more blurred boundary and are difficult to recognize can be delineated by MP-GAN. More specifically, the corpus callosum with irregular sulcus is depicted accurately by MP-GAN from the sagittal view and coronal view in Fig. 5. Atrophy of the corpus callosum may lead to functional disability because of reduced interhemispheric integration. It is a region that has been examined intensively for indications of EMCI [45]. On the other hand, integrated gradients and guided backprop tend to focus on some small parts of the lesions rather than the whole lesions. Because some subtle voxels of the lesion might be more salient than the other voxels of the whole lesion. This proves that the feature visualization methods based on classification only focus on the most discriminative features and ignore the rest. It is difficult to interpret the results produced by CAM due to low resolution. Moreover, the regions visualized by GAN seem to cover parts of ground truth affected by the AD for NC versus AD as shown in Fig. 7. However, they are not close to ground truth, and this

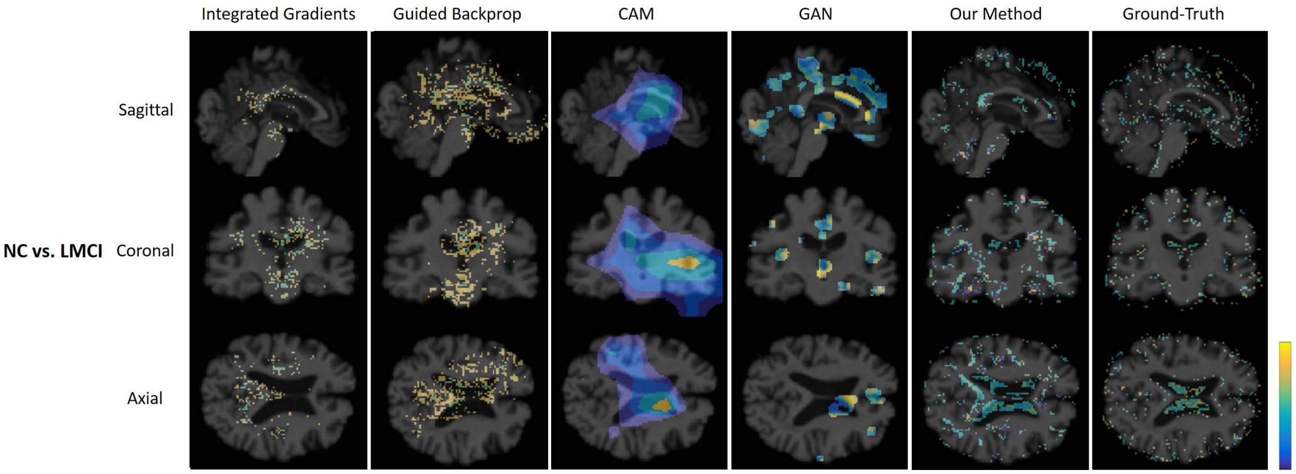


Fig. 6. Heatmaps predicted by integrated gradients, guided backprop, CAM, GAN, and our method are shown in sagittal, coronal, and axial views for NC versus LMCI, respectively.

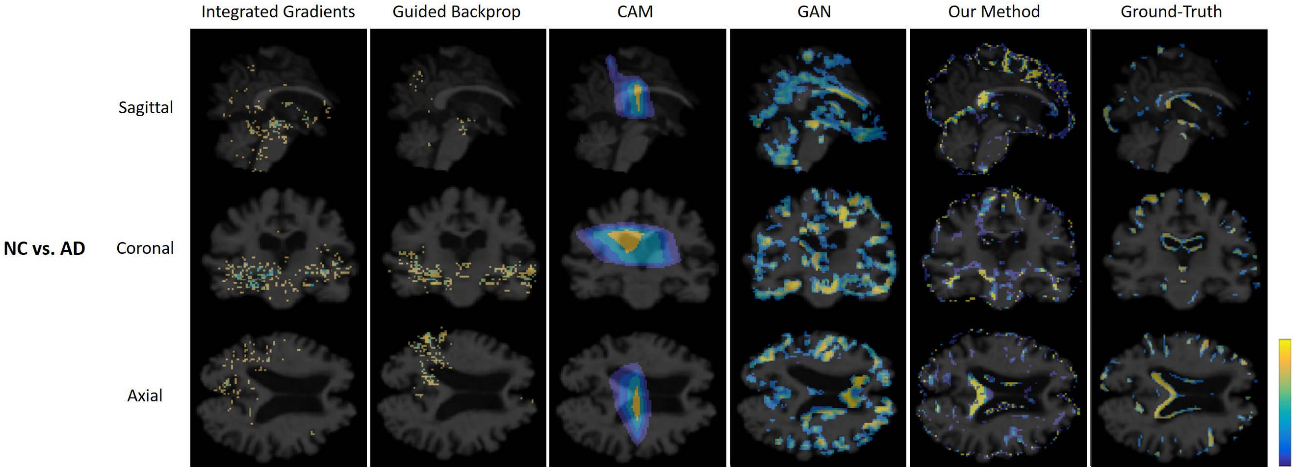


Fig. 7. Heatmaps predicted by integrated gradients, guided backprop, CAM, GAN, and our method are shown in sagittal, coronal, and axial views for NC versus AD, respectively.

is because the training of GAN is unstable. In summary, the results of the proposed MP-GAN are closer to the ground truth compared with the other four existing methods. This implies that MP-GAN can benefit from the multidirectional mapping mechanism and the hybrid loss function. MP-GAN is more sensitive to subtle structural changes in MR images caused by cognitive decline.

The ADNI diagnostic criteria for each Alzheimer's stage are briefly described as follows.

1) NC participants have no subjective or informant-based complaints of memory decline, and they have a normal cognitive performance.
2) SMC participants have subjective memory concerns assessed by cognitive change index (CCI). They have no informant-based complaint of memory impairment or decline, and they have a normal cognitive performance on Wechsler logical memory delayed recall (LM-delayed) and mini-mental state examination (MMSE) [46].
3) EMCI participants have a subtle cognitive decline. Their abnormal memory function is approximately 1 standard deviation below normative performance, and their MMSE total score is greater than 24.
4) LMCI participants have a memory concern. Clinical dementia rating (CDR) of LMCI participants is 0.5, and memory box (MB) score must be at least 0.5.
5) AD participants have an SMC. The MMSE score of AD participants is between 20 and 26, and CDR is 0.5 or 1.0.

To further analyze the visualization results of the proposed MP-GAN from a clinical perspective, the two-view slices in another coordinate of (33,55,39) are shown in Fig. 8. Note that the three-view slices shown from Figs. 4–7 are in the coordinate of (44,55,47). From Fig. 8, the following observations can be made.

1) For all the four evaluation groups, the proposed MP-GAN can delineate the discriminative lesions clearly. More specifically, lesions visualized by MP-GAN are hippocampus, thalamus, putamen, pallidum, caudate nucleus, amygdala, and insula [20], [47], [48]. It is worth noting that the discriminative capability of these brain regions in clinical analysis has already been validated by previous studies [1],

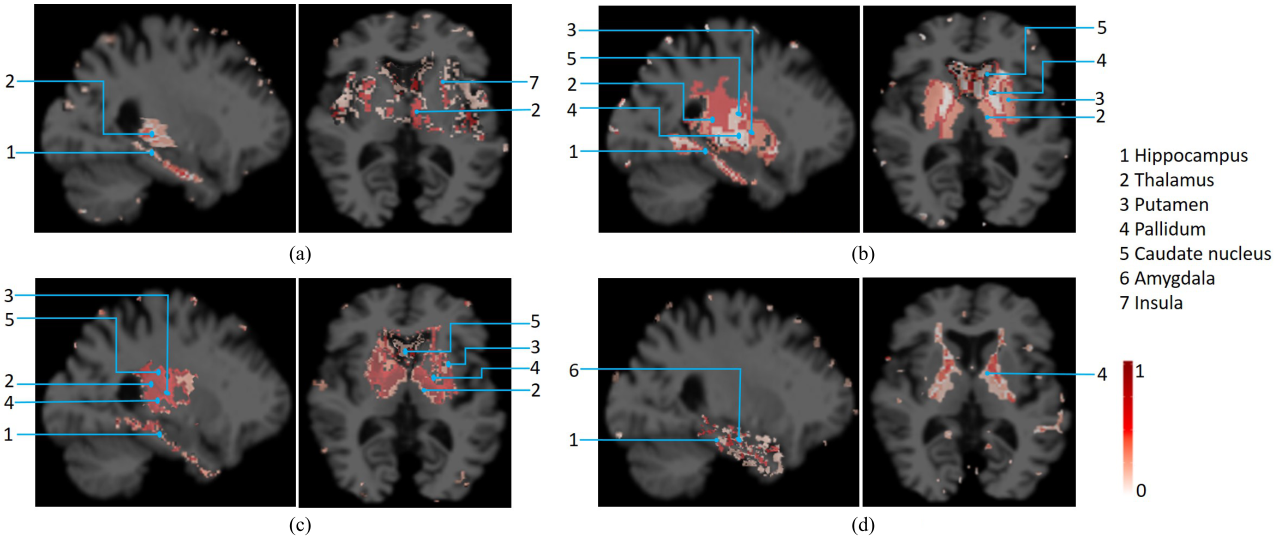


Fig. 8. Class discriminative maps generated by MP-GAN are shown as a colored overlay over MR images. The regions affected by progression of AD are reliably captured by MP-GAN for four evaluation groups, respectively. (a) NC versus EMCI. (b) NC versus LMCI. (c) NC versus AD. (d) EMCI versus AD.

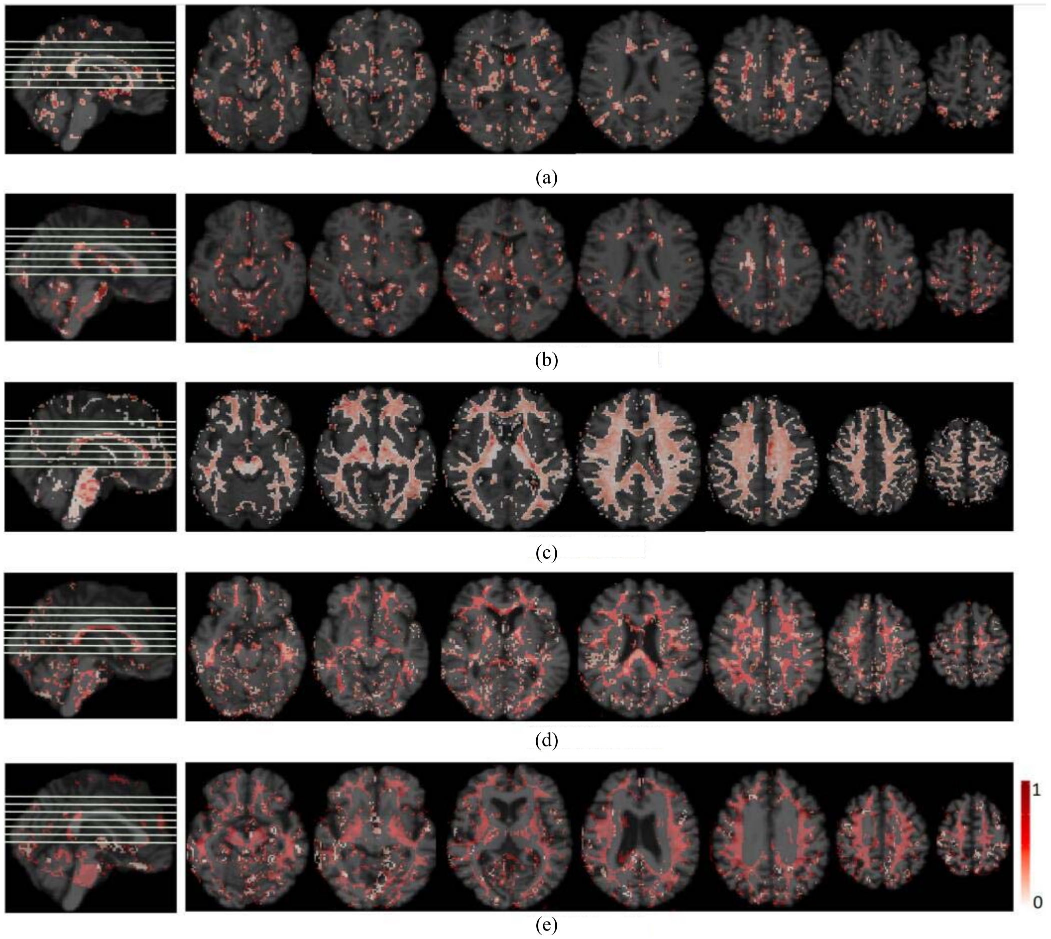
[49]–[51]. This implies the feasibility of the proposed MP-GAN.

2) The morphological changes including global atrophy (e.g., smaller volumes of hippocampus or amygdala) and shape changes are visualized by class discriminative map (indicated by color). These morphological changes are related to AD disease progression and cognitive decline severity.
3) For four evaluation groups, identified multiple regions are overlapped or localized at similar brain regions. For instance, the regions of NC versus LMCI and NC versus AD are similar because LMCI might develop to AD. Meanwhile, since the features between LMCI and AD are very subtle, some visualized regions of NC versus LMCI and NC versus AD are overlapped, but the atrophy severity of each lesion is different (indicated by color). The lesions visualized for EMCI versus AD and NC versus AD also have some common regions, such as the hippocampus and pallidum. Furthermore, it is reasonable that the overlap regions between NC versus EMCI and NC versus AD might not be identified for EMCI versus AD, and some regions such as the amygdala which are specific to EMCI versus AD can be identified.
4) Along with the progression from EMCI to AD, from Figs. 8(a)–(c), it can be observed that the intensity values (i.e., light salmon color) in the heatmaps are gradually increased (i.e., change to crimson) at various brain locations, and some of them are accumulated at the annotated regions.

These results suggest that the class discriminative maps generated by the proposed MP-GAN have the potential to provide some extra information regarding AD progression, and it may reveal the gradual atrophic process of human brain due to cognitive decline. Furthermore, the severity of cognitive decline is also reflected in ADNI diagnostic criteria for each Alzheimer's stage as described above. In summary, the above observations imply the robustness of MP-GAN in visualizing morphological features for different Alzheimer's stages.

For further visualization analysis, five evaluation groups are investigated, respectively, in Fig. 9. The results show that the important brain regions visualized by the proposed method are consistent with regions in Fig. 8. More specifically, by aligning the automatic anatomical labeling (AAL) map with the class discriminative maps visualized in Fig. 9, each region in the class discriminative map will be matched to the specific ROI index and name in AAL. The disease-related regions visualized by MP-GAN are listed in Table II. Note that the suffix "L" denotes the left brain, and the suffix "R" denotes the right brain. The following observations can be made from Fig. 9 and Table II: 1) the brain regions visualized by the proposed method for NC versus SMC are precental gyrus, middle frontal gyrus, inferior frontal gyrus, median cingulate, paracingulate gyri, parahippocampal gyrus, superior occipital gyrus, postcentral gyrus, and thalamus; 2) the brain regions visualized by the proposed method for SMC versus EMCI are rolandic operculum, insula, parahippocampal gyrus, amygdala, superior occipital gyrus, middle occipital gyrus, postcentral gyrus, superior parietal gyrus, and precuneus; 3) the brain regions visualized by the proposed method for NC versus EMCI are the middle frontal gyrus, posterior cingulate gyrus, calcarine fissure and surrounding cortex, cuneus, superior occipital gyrus, fusiform gyrus, postcentral gyrus, lenticular nucleus, putamen, and inferior temporal gyrus; 4) the brain regions visualized by the proposed method for EMCI versus LMCI are the superior frontal gyrus, orbital part, inferior frontal gyrus, opercular part, hippocampus, parahippocampal gyrus, calcarine fissure and surrounding cortex, lingual gyrus, inferior occipital gyrus, and fusiform gyrus; and 5) the brain regions visualized by the proposed method for LMCI versus AD are the middle frontal gyrus, orbital part, inferior frontal gyrus, triangular part, hippocampus, calcarine fissure and surrounding cortex, lingual gyrus, middle occipital gyrus, precuneus, lenticular nucleus, and putamen. These regions also agree with the existing research findings. To sum up, the lesions visualized by the proposed model are highly suggestive and effective for tracking the progression of AD.

The performance of MP-GAN to visualize the subtle lesions in the hippocampus is further investigated. The class discriminative maps of the hippocampus in the sagittal view are visualized in Fig. 10. Specifically, the following four neighborhood evaluation groups are further explored: 1) NC versus SMC;



(a)

(b)

(c)

(d)

(e)

Fig. 9. Distribution of class discriminative maps visualized by MP-GAN for five evaluation groups, respectively. (a) NC versus SMC. (b) SMC versus EMCI. (c) NC versus EMCI. (d) EMCI versus LMCI. (e) LMCI versus AD.

TABLE II

INDICES AND NAMES OF REGIONS VISUALIZED BY MP-GAN USING AAL TEMPLATE

NC vs. SMC		SMC vs. EMCI		NC vs. EMCI		EMCI vs. LMCI		LMCI vs. AD	
ROI index	ROI name	ROI index	ROI name	ROI index	ROI name	ROI index	ROI name	ROI index	ROI name
1	Precentral_L	17	Rolandic_Oper_L	7	Frontal_Mid_L	5	Frontal_Sup_Orb_L	9	Frontal_Mid_Orb_L
8	Frontal_Mid_R	29	Insula_L	36	Cingulum_Post_R	12	Frontal_Inf_Oper_R	14	Frontal_Inf_Tri_R
10	Frontal_Mid_Orb_R	39	ParaHippocampal_L	43	Calcarine_L	37	Hippocampus_L	37	Hippocampus_L
12	Frontal_Inf_Oper_R	42	Amygdala_R	45	Cuneus_L	39	ParaHippocampal_L	38	Hippocampus_R
34	Cingulum_Mid_R	49	Occipital_Sup_L	46	Cuneus_R,	40	ParaHippocampal_R	43	Calcarine_L
39	ParaHippocampal_L	51	Occipital_Mid_L	50	Occipital_Sup_R	43	Calcarine_L	48	Lingual_R
40	Parahippocampal_R	52	Occipital_Mid_R	56	Fusiform_R	47	Lingual_L	52	Occipital_Mid_R
50	Occipital_Sup_R	58	Postcentral_R	57	Postcentral_L	50	Occipital_Sup_R	67	Precuneus_L
57	Postcentral_L	60	Parietal_sup_R	74	Putamen_R	54	Occipital_Inf_R	68	Precuneus_R
78	Thalamus_R	67	Precuneus_L	90	Temporal_Inf_R	55	Fusiform_L	74	Putamen_R

2) SMC versus EMCI; 3) EMCI versus LMCI; and 4) LMCI versus AD. From Fig. 10, it can be observed that the zoomed regions preserve more details in the hippocampus. In particular, in the earlier stages of AD such as: 1) NC versus SMC and 2) SMC versus EMCI, the visualized lesions are extremely subtle and scattered around the boundary of the hippocampus. In the later stages of AD such as: 1) EMCI versus LMCI and 2) LMCI versus AD, the visualized lesions are accumulated at the core region of the hippocampus. Furthermore, Fig. 10(a)–(d) reflects the shape change and atrophy of the hippocampus qualitatively as the progressive deterioration from SMC to AD. It has already been validated by previous studies [52] that the hippocampus is significant for identifying biomarkers in clinical practice. Although the volume loss and shape change of the hippocampus cannot be quantitatively measured in this work, the visualized lesions of the hippocampus are beneficial for identifying the biomarkers in future work. Based on these visualized lesions in Fig. 10, the existing biomarkers such as brain boundary shift integral (BBSI) [53], scoring by nonlocal image patch estimator (SNIPE) [54], and other grading biomarkers [55] can be computed. Furthermore, new potential biomarkers reflecting the shape change and brain atrophy might be discovered based on these visualized lesions in the hippocampus in future work.

D. Quantitative Analysis

In this section, the following four metrics are computed to assess visual quality.

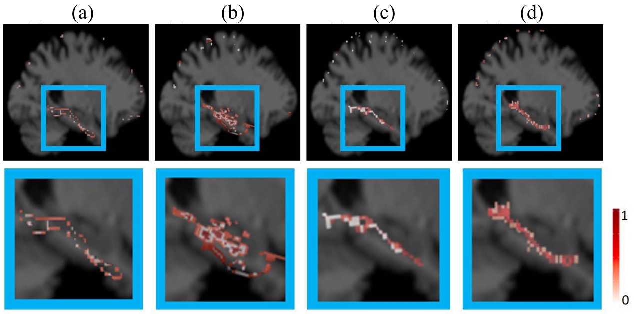


Fig. 10. Visualization results of hippocampus by MP-GAN in sagittal view and corresponding zoomed regions. The subfigures at the bottom are the zoom of the original subfigures for better observation. (a) NC vs. SMC. (b) SMC vs. EMCI. (c) EMCI vs. LMCI. (d) LMCI vs. AD.

1) *Normalized Cross Correlation (NCC):* NCC [44] is calculated between ground-truth maps and predicted class discriminative maps. The higher the NCC, the more correlation between ground truth maps and the predicted class discriminative maps. For integrated gradients, guided backprop, and CAM, the visualized heatmaps for predicting positive class are used to calculate the NCC.
2) *Peak Signal-to-Noise Ratio (PSNR):* PSNR [56] is also calculated between ground-truth maps and predicted class discriminative maps on the test dataset. Similar to NCC, the higher the PSNR, the closer the ground-truth maps and the predicted class discriminative maps.
3) *Structural Similarity Index Measure (SSIM) [57]:* Different from NCC and PSNR, SSIM in each iteration is calculated between synthetic images and real images on the validation dataset. Higher SSIM indicates better reconstructed MR image quality. By computing SSIM in each iteration, the convergency of the model is further validated.
4) *Classification Metrics [4] Such as AUC, ACC, Sensitivity, and Specificity for Data Augmentation:* Note that the purpose of SSIM and classification metrics is to demonstrate that the proposed MP-GAN can generate images close to real distribution, and thus it validates that MP-GAN can capture salient global features in class discriminative maps.

For NCC and PSNR, the four existing methods are compared. For SSIM, only GAN is compared since the other three methods are based on classification. Similarly, for classification metrics, only GAN is compared since the classification performance is based on synthetic data augmentation by the proposed MP-GAN and GAN.

The NCC results shown in Fig. 11 are mostly consistent with the qualitative results shown from Figs. 4–7. The proposed MP-GAN achieves significantly higher NCC than the other four existing methods. It indicates that the distribution of class discriminative maps generated by MP-GAN is the closest to ground-truth maps. The three methods based on classification (integrated gradients, guided backprop, and CAM) achieve a low NCC score due to their exclusive focus on local features. GAN performs better than three classification-based feature visualization methods for NC versus SMC, NC versus LMCI, and NC versus AD. This implies that the GAN architecture can capture global features, which alleviates the limitations of feature visualization methods based on classification. Above all, the proposed MP-GAN achieves the highest correlation scores compared with the other four existing methods in all the four evaluation groups.

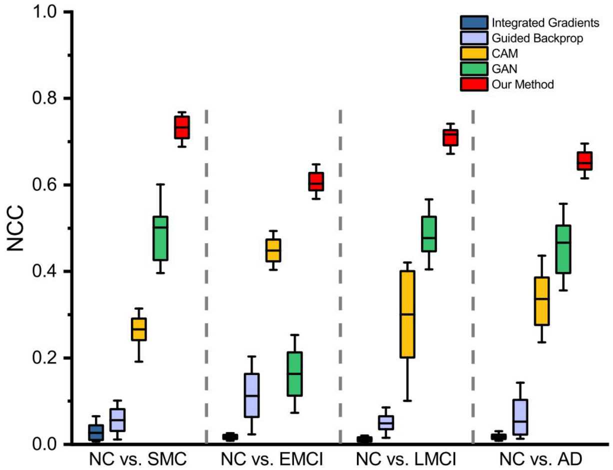


Fig. 11. Box plots of NCC for different models.

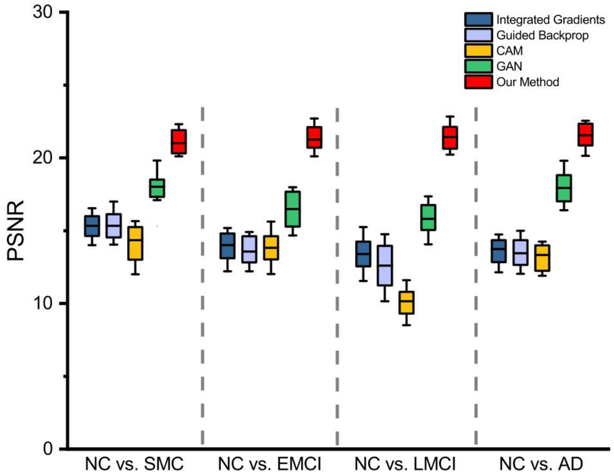


Fig. 12. Box plots of PSNR for different models.

From Fig. 12, it can be seen that the proposed MP-GAN achieves the best PSNR compared with the other four existing methods. This is also consistent with NCC results in Fig. 11 and the qualitative results shown from Figs. 4–7. The class discriminative maps visualized by MP-GAN are closer to ground truth. This is because MP-GAN benefits from the multidirectional mapping mechanism and hybrid loss function. Meanwhile, MP-GAN can be trained on MR images of all the classes with only one model. In this manner, the common features unrelated to the disease can be reused, and thus all the salient global features can be captured in class discriminative maps for different Alzheimer's stages.

Furthermore, the generation diversity with SSIM is evaluated in each iteration on the validation dataset. The convergence curves of the proposed MP-GAN and GAN are given for four evaluation groups: 1) NC versus SMC; 2) NC versus EMCI; 3) NC versus LMCI; and 4) NC versus AD, respectively. From Figs. 13–16, it can be observed that the proposed MP-GAN converges faster than GAN. Meanwhile, MP-GAN performs stably in all four evaluation groups. On the other hand, the training of GAN is extremely unstable for NC versus LMCI, and it cannot converge for NC versus EMCI

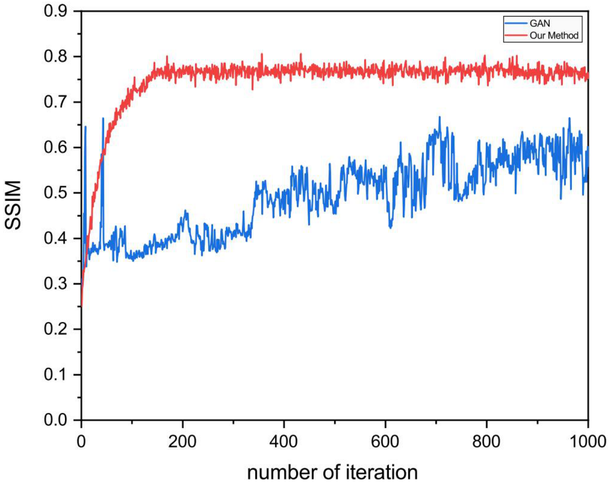


Fig. 13. Convergence curves for NC versus SMC.

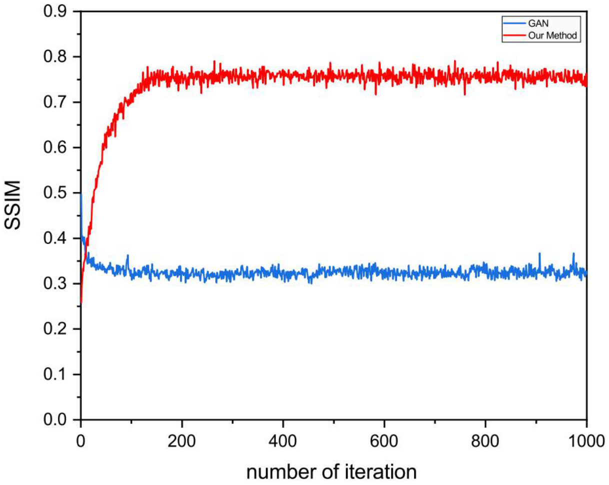


Fig. 14. Convergence curves for NC versus EMCI.

and NC versus AD. Again, these results are consistent with the NCC score in Fig. 11. The NCC score of GAN is low for NC versus EMCI in Fig. 11, because GAN cannot converge for NC versus EMCI as shown in Fig. 14. These results also indicate that the proposed MP-GAN can generate diverse MR images close to real distribution.

The objective of synthetic data augmentation is to demonstrate the learned class discriminative map have captured all the subtle morphological features for different stages of AD. If the classification performance is improved, this indicates that the learned class discriminative maps have captured all the subtle morphological features for different stages of AD progression. Thus, the synthetic MR images produced by class discriminative maps can be classified as the corresponding class correctly. By conducting this experiment, the efficacy of MP-GAN is corroborated. More specifically, the CNN classifier is trained using synthetic data augmentation. More specifically, the 100 synthesized MR images of each class by MP-GAN and GAN are added to the original training set to form two new augmented training sets separately. Then the CNN model is trained on the two new augmented training sets separately for each evaluation group. During the test stage, the same test set of real MR images is used. From Figs. 17–20, it can be seen that adding synthesized samples by the proposed MP-GAN achieves better classification performance in terms of AUC, accuracy, specificity, and sensitivity. Overall, the synthetic data samples generated by MP-GAN can add additional

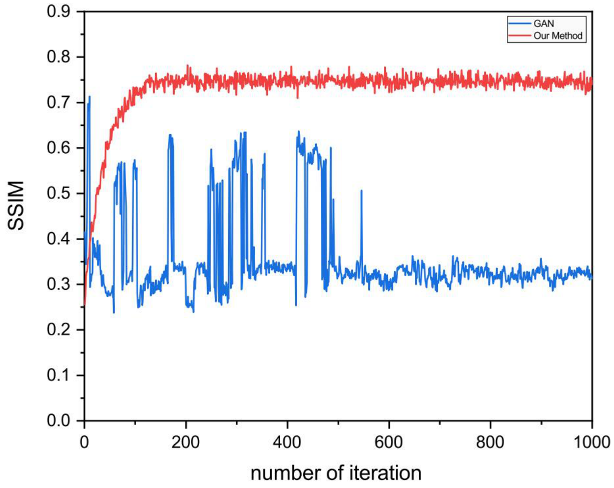


Fig. 15. Convergence curves for NC versus LMCI.

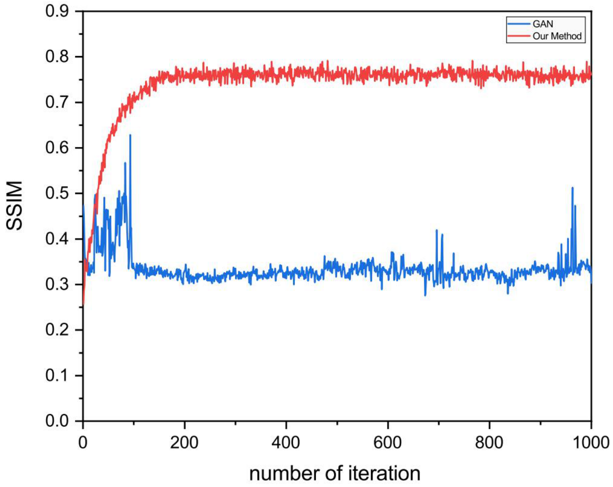


Fig. 16. Convergence curves for NC versus AD.

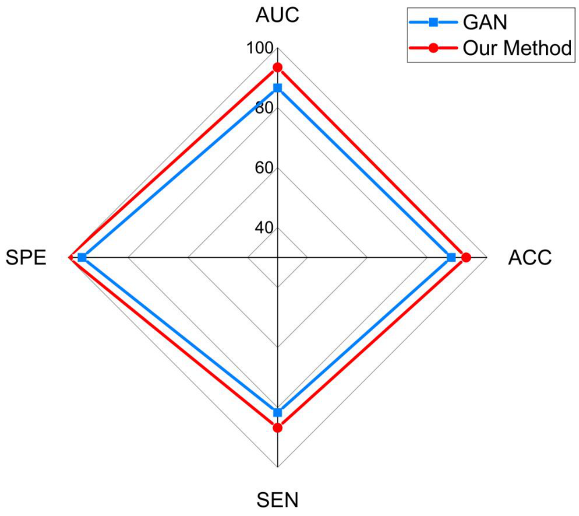


Fig. 17. Classification results of synthetic data augmentation for NC versus SMC.

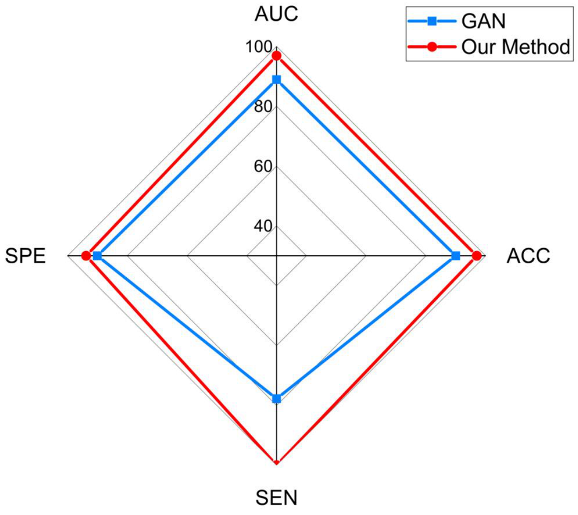


Fig. 18. Classification results of synthetic data augmentation for NC versus EMCI.

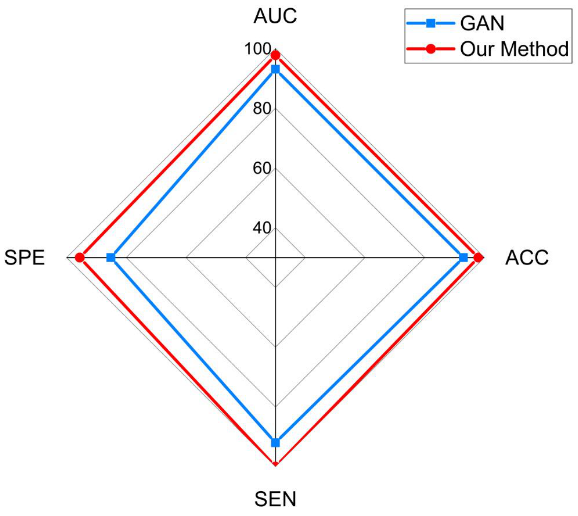


Fig. 19. Classification results of synthetic data augmentation for NC versus LMCI.

variability to the original training set, which in turn leads to better performance. This implies that the synthesized MR images generated by MP-GAN not only provide meaningful visualizations but also capture discriminative features for AD analysis. The proposed MP-GAN can be used as an effective data augmentation method.

V. Discussion

Although extensive experiments demonstrate the superiority of the proposed MP-GAN, MP-GAN has two limitations.

1) The hyperparameters are tuned empirically for the best performance. The optimal value of the hyperparameters depends on network architecture and data. There is no straightforward way to find optimal hyperparameters in advance.
2) The conventional GAN has several common failure modes, such as training instability and mode collapse. The proposed MP-GAN is designed as a three-player cooperative game instead of the conventional two-player competition game by introducing the auxiliary classifier network based on the generator and the discriminator.

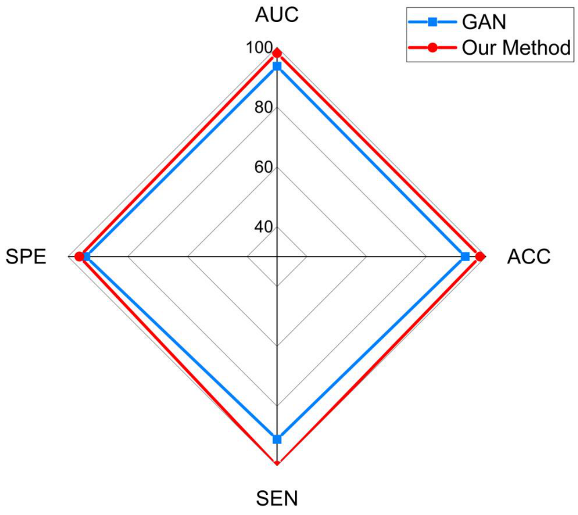


Fig. 20. Classification results of synthetic data augmentation for NC versus AD.

The specific architecture design and the proposed hybrid loss can make the training process more stable. However, mode collapse, which is a common issue of the GAN model, still might happen even when MP-GAN has shown stable training performance. In future works, how to solve the mode collapse issue is the direction to further improve the robustness of the proposed MP-GAN.

VI. Conclusion

In this article, a novel MP-GAN is proposed to visualize the morphological features indicating the severity of AD in whole-brain MR images. By introducing a novel multidirectional mapping mechanism into the model, MP-GAN can capture the salient global features efficiently. Thus, using the class-discriminative map from the generator, the proposed model can clearly delineate the subtle lesions via MR image transformations between the source domain and the target domain. Besides, by integrating the adversarial loss, classification loss, cycle consistency loss, and $L1$ penalty, a single generator in MP-GAN can learn the class-discriminative maps for multiple classes. The experimental results on the public ADNI dataset have demonstrated that MP-GAN can visualize multiple lesions affected by the progression of AD accurately. Furthermore, MP-GAN may visualize some new disease-related regions that have not been investigated yet. This can be studied further to discover potential new AD biomarkers in future work.

Acknowledgment

The authors would like to thank the data support of ADNI.

References

[1] X. Hao *et al.*, “Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer’s disease,” *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101625.

[2] M. W. Bondi, E. C. Edmonds, and D. P. Salmon, “Alzheimer’s disease: Past, present, and future,” *J. Int. Neuropsychol. Soc.*, vol. 23, nos. 9–10, pp. 818–831, Oct. 2017.

[3] Y. Shi *et al.*, "Leveraging coupled interaction for multimodal Alzheimer's disease diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 186–200, Jan. 2019.

[4] S. Wang, H. Wang, Y. Shen, and X. Wang, "Automatic recognition of mild cognitive impairment and Alzheimers disease using ensemble based 3D densely connected convolutional networks," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 517–523.

[5] S. Wang, Y. Shen, W. Chen, T. Xiao, and J. Hu, "Automatic recognition of mild cognitive impairment from MRI images using expedited convolutional neural networks," in *Proc. Int. Conf. Artif. Neural Netw.*, 2017, pp. 373–380.

[6] N. Mammone, C. Ieracitano, H. Adeli, A. Bramanti, and F. C. Morabito, "Permutation Jaccard distance-based hierarchical clustering to estimate EEG network density modifications in MCI subjects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 5122–5135, Oct. 2018.

[7] M. Wang, D. Zhang, D. Shen, and M. Liu, "Multi-task exclusive relationship learning for Alzheimer's disease progression prediction with longitudinal data," *Med. Image Anal.*, vol. 53, pp. 111–122, Feb. 2019.

[8] B. Jie, M. Liu, and D. Shen, "Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease," *Med. Image Anal.*, vol. 47, pp. 81–94, Jul. 2018.

[9] G. S. Babu and S. Suresh, "Sequential projection-based metacognitive learning in a radial basis function network for classification problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 194–206, Feb. 2013.

[10] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.

[11] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," 2018, *arXiv:1809.07294*.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[14] W. Yu, B. Lei, M. K. Ng, A. C. Cheung, Y. Shen, and S. Wang, "Tensorizing GAN with high-order pooling for Alzheimer's disease assessment," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 17, 2021, doi: 10.1109/TNNLS.2021.3063516.

[15] S. Wang *et al.*, "Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 574–585, Apr. 2021.

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 70, D. Precup and Y. W. Teh, Eds. PMLR, Aug. 2017, pp. 214–223.

[17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5769–5779.

[18] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, "Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease," 2020, *arXiv:2004.12204*.

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2014, pp. 818–833.

[20] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3D brain MRI classification," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 835–838.

[21] A. Mahendran and A. Vedaldi, "Salient deconvolutional networks," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2016, pp. 120–135.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR*, 2014, pp. 1–8.

[23] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[25] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.

[26] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Frontiers Aging Neurosci.*, vol. 11, p. 194, Jul. 2019.

[27] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.

[28] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017, *arXiv:1703.01365*.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[30] N. M. Khan, N. Abraham, and M. Hon, "Transfer learning with intelligent training data selection for prediction of Alzheimer's disease," *IEEE Access*, vol. 7, pp. 72726–72735, 2019.

[31] C. Lian, M. Liu, L. Wang, and D. Shen, "End-to-end dementia status prediction from brain MRI using multi-task weakly-supervised attention network," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, Jan. 2019, pp. 158–167.

[32] S. Sarraf and G. Tofighi, "Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks," 2016, *arXiv:1607.06583*.

[33] T. Kim, M. Cha, H. Kim, J. Kwon Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," 2017, *arXiv:1703.05192*.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016.

[35] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[37] D. Gu, "3D densely connected convolutional network for the recognition of human shopping actions," School Elect. Eng. Comput. Sci., Dept. Eng., Univ. Ottawa, Ottawa, ON, Canada, Tech. Rep., 2017. [Online]. Available: http://hdl.handle.net/10393/36739, doi: 10.20381/ruor-21013.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[39] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *NeuroImage*, vol. 45, no. 1, pp. S173–S186, Mar. 2009.

[40] S. M. Smith *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, pp. S208–S219, Jan. 2004.

[41] S. M. Smith, "Fast robust automated brain extraction," *Hum. Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[42] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, Jun. 2001.

[43] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, Oct. 2002.

[44] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using Wasserstein GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8309–8319.

[45] B. A. Ardekani, A. H. Bachman, K. Figarsky, and J. J. Sidtis, "Corpus callosum shape changes in early Alzheimer's disease: An MRI study using the OASIS brain database," *Brain Struct. Function*, vol. 219, no. 1, pp. 343–352, 2014.

[46] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, and X. Li, "Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1508–1519, Jul. 2017.

[47] Y. Zhang *et al.*, "Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning," *Frontiers Comput. Neurosci.*, vol. 9, p. 66, Jun. 2015.

[48] J. M. Rondina *et al.*, "Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases," *NeuroImage, Clin.*, vol. 17, pp. 628–641, Jan. 2018.

[49] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, and B. Lei, “Deep learning framework for Alzheimer’s disease diagnosis via 3D-CNN and FSBi-LSTM,” *IEEE Access*, vol. 7, pp. 63605–63618, 2019.

[50] H. Braak and E. Braak, “Neuropathological stageing of Alzheimer-related changes,” *Acta Neuropathol.*, vol. 82, no. 4, pp. 239–259, 1991.

[51] B. C. Dickerson *et al.*, “The cortical signature of Alzheimer’s disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals,” *Cerebral Cortex*, vol. 19, no. 3, pp. 497–510, Mar. 2009.

[52] F. Mrquez and M. A. Yassa, “Neuroimaging biomarkers for Alzheimer’s disease,” *Mol. Neurodegeneration*, vol. 14, no. 1, p. 21, Jun. 2019.

[53] N. C. Fox and P. A. Freeborough, “Brain atrophy progression measured from registered serial MRI: Validation and application to alzheimer’s disease,” *J. Magn. Reson. Imag.*, vol. 7, no. 6, pp. 1069–1075, Nov. 1997.

[54] P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, and D. L. Collins, “Simultaneous segmentation and grading of anatomical structures for patient’s classification: Application to Alzheimer’s disease,” *NeuroImage*, vol. 59, no. 4, pp. 3736–3747, 2012.

[55] T. Tong, Q. Gao, R. Guerrero, C. Ledig, L. Chen, and D. Rueckert, “A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer’s disease,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 155–165, Jan. 2017.

[56] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

[57] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2004, pp. 1398–1402.

Wen Yu received the Ph.D. degree from the Department of Computer Science, University of Liverpool, Liverpool, U.K., in 2015.

She was a Senior Software Engineer at Hong Kong and Shanghai Banking Corporation Ltd. (HSBC) from 2016 to 2018. She held a Post-Doctoral Fellowship at the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China, from 2018 to 2021. Her research interests include deep learning and computer vision.

Baiying Lei (Senior Member, IEEE) received the M.Eng. degree in electronics science and technology from Zhejiang University, Hangzhou, China, in 2007, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2013.

She is currently with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen, China. She has coauthored more than 180 scientific articles, e.g., *Medical Image Analysis*, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Cybernetics (TCYB), IEEE Transactions on Medical Imaging (TMI), and IEEE Transactions on Biomedical Engineering (TBME). Her research interests include medical image analysis, machine learning, and pattern recognition.

Dr. Lei serves as an Associate Editor for the IEEE TMI and an Editorial Board Member of *Medical Image Analysis*, *Neural Computing and Application*, *Frontiers in Neuroinformatics*, *Frontiers in Aging Neuroscience?Scientific Reports*, and *PLOS One*.

Shuqiang Wang (Member, IEEE) received the Ph.D. degree in system engineering and engineering management from the City University of Hong Kong, Hong Kong, in 2012.

He was a Research Scientist of the Noah’s Ark Laboratory, Huawei Technologies, Shenzhen, China. He held a Post-Doctoral Fellowship at The University of Hong Kong, Hong Kong, from 2013 to 2014. He is currently a Professor with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen. His research interests include machine learning, medical image computing, and optimization theory.

Yong Liu was born in 1986. He received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016.

He is currently an Associate Professor with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. His research interests include large-scale machine learning, large-scale model selection, and auto machine learning.

Zhiguang Feng received the Ph.D. degree from the Department of Mechanical Engineering, The University of Hong Kong, Hong Kong, in 2013.

In 2019, he was a Vice-Chancellor’s Post-Doctoral Research Fellow at the University of Wollongong, Wollongong, NSW, Australia. He is currently a Professor at the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China. His research interests include singular systems, time-delay systems, robust control, dissipative control, and reachable set estimation.

Yong Hu (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in biomedical engineering from Tianjin University, Tianjin, China, in 1985 and 1988, respectively, and the Ph.D. degree from The University of Hong Kong, Hong Kong, in 1999.

He is currently an Associate Professor and the Director of the Neural Engineering and Clinical Electrophysiology Laboratory, Department of Orthopedics and Traumatology, The University of Hong Kong. His research interests include neural engineering, clinical electrophysiology, and biomedical signal measurement and processing.

Yanyan Shen (Member, IEEE) received the Ph.D. degree from the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong, in 2012.

From 2013 to 2014, she was a Post-Doctoral Research Fellow with the School of Information and Communication Engineering, Inha University, Incheon, South Korea. She is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include optimization methods and machine learning in wireless networks.

Michael K. Ng (Senior Member, IEEE) received the B.Sc. and M.Phil. degrees from The University of Hong Kong, Hong Kong, in 1990 and 1992, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 1995.

He is currently a Chair Professor with the Department of Mathematics, The University of Hong Kong. His research interests include data science, imaging science, and scientific computing.

Dr. Ng also serves on the editorial boards of international journals (see https: //hkumath.hku.hk/ mng/).