Learn-Explain-Reinforce: Counterfactual Reasoning and Its Guidance to Reinforce an Alzheimer's Disease Diagnosis Model

Kwanseok Oh*, Jee Seok Yoon*, and Heung-Il Suk, *Senior Member, IEEE*

Abstract—Existing studies on disease diagnostic models focus either on diagnostic model learning for performance improvement or on the visual explanation of a trained diagnostic model. We propose a novel learn-explain-reinforce (LEAR) framework that unifies diagnostic model learning, visual explanation generation (explanation unit), and trained diagnostic model reinforcement (reinforcement unit) guided by the visual explanation. For the visual explanation, we generate a counterfactual map that transforms an input sample to be identified as an intended target label. For example, a counterfactual map can localize hypothetical abnormalities within a normal brain image that may cause it to be diagnosed with Alzheimer's disease (AD). We believe that the generated counterfactual maps represent data-driven and model-induced knowledge about a target task, *i.e.*, AD diagnosis using structural MRI, which can be a vital source of information to reinforce the generalization of the trained diagnostic model. To this end, we devise an attention-based feature refinement module with the guidance of the counterfactual maps. The explanation and reinforcement units are reciprocal and can be operated iteratively. Our proposed approach was validated via qualitative and quantitative analysis on the ADNI dataset. Its comprehensibility and fidelity were demonstrated through ablation studies and comparisons with existing methods.

Index Terms—Visual Explanation, Counterfactual Reasoning, Representation Reinforcement, Explanation-Guided Attention, Deep Learning, Explainable AI (XAI), Structural Magnetic Resonance Imaging, Alzheimer's Disease.

1 Introduction

Alzheimer's disease (AD) is known as one of the most prevalent neurodegenerative diseases, characterized by progressive and irreversible memory loss and cognitive function decline or impairment. AD causes the damage and destruction of nerve cells in brain regions related to memory, language, and other cognitive functions, and it has contributed to 60–80% of the world's dementia cases [1]. Brain atrophy associated with AD emerges as a *continuous* progression from cognitively normal (CN) to mild cognitive impairment (MCI) and dementia in the symptomatic spectrum [2]. Currently available AD-related medicines have marginal effects in alleviating amnesic symptoms or slowing their progression. Thus, early detection and timely intervention of AD at its preclinical or prodromal stages are of paramount importance in the prevention of its progression and in diminishing its incidence.

Of various brain imaging tools, structural magnetic resonance imaging (sMRI) has been most intensively studied for AD diagnosis as it provides imaging biomarkers of neuronal loss in the anatomical structures of a brain [3]. Specifically, sMRI scans are helpful in detecting and measuring morphological changes in the brain, such as enlarged ventricle and regional atrophies, and anatomical variations across subjects. In the last few decades, researchers have devoted their efforts to devising machine-learning techniques that can analyze and identify the potential risk of a subject having AD or MCI at an early stage [4], [5], [6], [7], [8]. More recently, with the unprecedented advances in deep learning, there have been many successful studies in sMRI-based AD diagnosis that achieved clinically applicable performance [9], [10], [11], [12], [13].

In the meantime, there has been a growing need for explainability of a model's output and/or interpretability of a model's internal workings [14], [15], [16]. The black-box nature of deep learning models limits their real-world application in the fields of medicine, security, and finance, especially where fairness, accountability, and transparency are essential. From the end-user's (*e.g.*, clinicians and patients) point of view, it is crucial to be able to interpret and explain a deep-learning model's output at the level of human knowledge and understanding. However, building a predictive model for high performance that is also equipped with interpretability or explainability is still an unsolved problem because of their trade-off, *i.e.*, interpretable/explainable models tend to have lower performance than black-box models [17], [18], [19], especially in the field of medical vision.

Reducing this trade-off between performance and interpretability/explainability has been a long-standing goal in the field of explainable AI (XAI). In the early era of XAI [20], researchers have proposed various methods for discovering or identifying the regions that have the most influence on deriving the outcome of a classifier [21], [22], [23], [24], [25], [26]. The main objective of those XAI methods is to

- *K. Oh is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: ksohh@korea.ac.kr).*
- *J. Yoon is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: wltjr1007@korea.ac.kr).*
- *H.-I. Suk is with the Department of Artificial Intelligence and the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea and the corresponding author (e-mail: hisuk@korea.ac.kr).*
- ** indicates equal contribution.*

answer the question, *"For an input X, which part influenced the classifier's decision to label it Y?"* However, recent XAI methods try to answer the question that can offer a more fundamental explanation: *"If an input X was X^*, would the outcome have been Z rather than Y?"* [27], [28], [29] in the sense of causality. This sort of explanation is defined at the root of *counterfactual reasoning* [30]. Counterfactual reasoning can provide an explanation at the level of human knowledge as it explains a model's decision in hypothetical scenarios.

Inspired by this philosophical concept of counterfactual reasoning, in this work, we propose a novel method for a higher-level visual explanation of a deep predictive model designed and trained for AD diagnosis using sMRI. Specifically, our method generates a '*counterfactual map*' conditioned on a target label (*i.e.*, hypothetical scenario). This map is added to the input image to transform it to be diagnosed as a target label. For example, when a counterfactual map is added to the input MRI of AD subject, it causes the input MRI to be transformed such that it will be diagnosed as CN [31], [32]. Most of the existing works on producing a counterfactual explanation exploit generative models with generative adversarial network (GAN) and its variants [33], [34], [35]. To the best of our knowledge, however, they are limited to producing a single-way [28], [33], [36] or dual-way [27], [29] explanation. In other words, they only consider one or two hypothetical scenarios for counterfactual reasoning (*e.g.*, single-way counterfactual map can only transform a CN subject to an AD patient, and vice versa for dual-way maps). Thus, when there are more than two classes of interest for diagnosis, *e.g.*, CN *vs.* MCI *vs.* AD, a set of such explainable models must be built separately and independently for different pairs of clinical labels, *e.g.*, CN *vs.* MCI, MCI *vs.* AD, and CN *vs.* AD. However, with those separately and independently trained explanation models, it is likely to be incompatible and inconsistent in explanation, especially, in terms of the AD spectrum, raising accountability or interpretability issues. Consequently, it is necessary to build a single model for multi-way counterfactual map generation. Notably, a multi-way counterfactual map for an AD diagnostic model can provide a natural proxy for the stratification of diseases by producing hypothetical scenarios for intermediate stages (*e.g.*, CN→MCI→AD) of a disease. To this end, we propose a novel *multi-way* counterfactual reasoning method such that we can produce counterfactual maps for transforming an input to be any of the clinical labels under consideration (*i.e.*, CN, MCI, and AD).

Meanwhile, we believe it is desirable to utilize the counterfactual maps as *privileged information*, derived from an explanation model in combination with an AD diagnostic model during a training stage, to further enhance a classifier's generalizability, thus improving performance. In particular, thanks to the favorable counterfactual map's localization properties, we propose to exploit such information to guide a diagnostic model's focus on learning representations and discovering disease-related discriminative regions, which can be regarded as anatomical landmarks for diagnosis.

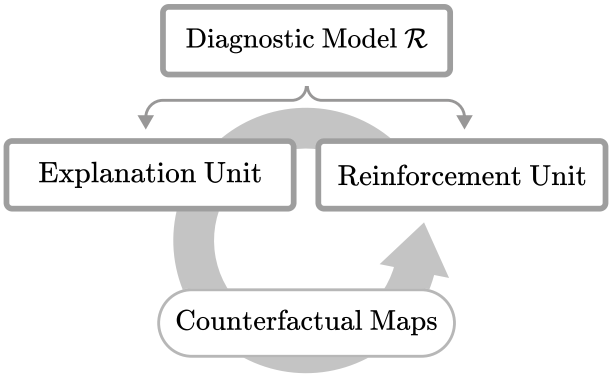

Fig. 1: Schematic of our proposed learn-explain-reinforce (LEAR)[†] framework. The explanation unit is a variation of conditional GAN that can synthesize a counterfactual map conditioned on an arbitrary target label. The reinforcement unit provides adequate guidance from the produced counterfactual map for reinforcing the performance of diagnostic model $\mathcal{R}$. We also introduce a simple iterative optimization scheme that enables simultaneous improvement of the explanation and diagnostic performance.

To this end, we propose a novel learn-explain-reinforce (LEAR)[†] framework. Our LEAR framework can produce high-quality counterfactual maps with state-of-the-art diagnostic performances through explanation-guided model reinforcement. Fig. 1 illustrates the schematic diagram of our proposed framework for counterfactual map generation to explain a diagnostic model's output (Explanation Unit) and its use to reinforce the generalizability of the diagnostic model via our newly devised pluggable reinforcement unit.

The main contributions of our work can be summarized as follows:

- We propose a novel learn-explain-reinforce framework that integrates the following tasks: (1) training a diagnostic model, (2) explaining a diagnostic model's output, and (3) reinforcing the diagnostic model based on the explanation systematically. To the best of our knowledge, this work is the first that exploits an explanation output to improve the generalization of a diagnostic model reciprocally.
- In regard to explanation, we propose a GAN-based method to produce *multi-way* counterfactual maps that can provide a more precise explanation, accounting for severity and/or progression of AD.
- Our work qualitatively and quantitatively surpasses state-of-the-art works in visual explanation and classification performance simultaneously.

The remainder of this article is organized as follows. In Section 2, we briefly review related work on attribution-based approaches and counterfactual explanations. Next, we introduce the automated AD diagnosis using attention

[†] The term reinforce/reinforcement used in this article refer to reinforcing the visual explanation and diagnostic models by means of attention-based guidance. Thus, it should *not* be confused with reinforcement learning in machine learning, which is a learning paradigm for intelligent agents.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3197845

TABLE 1: Recent studies categorized into visual explanation, use of attention mechanism, explanation-guided methods, and ability to reinforce visual explanation.

Methods	Visual Explanation	Attention	Guidance	Reinforce	Description
Liu *et al.* [37]	LRP [23]				Improving the diagnostic performance through instance normalization and model capacity increase
Korolev *et al.* [10]	-				Unique feature extraction by applying the dropout operation before the fully connected layer
Jin *et al.* [38]	-	✓			Discriminative feature extraction using the attention-based residual network
Zhang *et al.* [39]	Grad-CAM [21]	✓			Global and local representation captured using self-attention with the residual connection
Lian *et al.* [40]	CAM [41]	✓	✓		Attention-guided anatomical landmarks to capture multi-level discriminative patches and regions
Li *et al.* [42]	CAM [41]		✓	✓	Iterative attention-focusing strategy for joint pathological region localization and identification
LEAR (Ours)	Counterfactual Reasoning [30]	✓	✓	✓	Reinforcement of the diagnostic performance and explainability via the self-iterative training strategy with guidance

with guidance. The proposed method is described in detail in Section 3. In Section 4, we describe the studied datasets (*i.e.*, ADNI-1 and ADNI-2) with the data preprocessing pipeline as well as the experimental settings, competing methods, and qualitative and quantitative experimental results. We conclude this article and briefly discuss our stance on model explanation in Section 5. Our code is available at: https://github.com/ku-milab/LEAR.

2 Related Work

In this section, we describe various works proposed for explainable AI (XAI) and attention with guidance approaches for the improvement of AD diagnosis using sMRI.

2.1 Attribution-based Explanations

Attribution-based explanation refers to discovering or identifying the regions that have the most influence on deriving the outcome of a model. The methodological approaches for attribution-based explanation can be subdivided into gradient-based methods and reference-based methods. A gradient-based method highlights the activation nodes that contributed the most to a model's output. For example, class activation map (CAM) [41], and Grad-CAM [21] highlight the activation patterns of weights in a specified layer. Similarly, DeepTaylor [24], DeepLift [25], and layer-wise relevance propagation (LRP) [23] pinpoint the attributes that contributed to a model's output score by tracing back via gradient-based computations. These methods usually suffer from vanishing gradients especially when using ReLU activation. Integrated Gradients [22] resolves this issue through sensitivity analysis. Note that, gradient-based methods fundamentally explain the output decision based on the discriminative abstract features at the upper layers close to the classifier. Due to the lack of localization information in the coarse high-level feature maps, attribution-based methods [21], [22], [23], [24], [41] mostly suffer from providing blurry saliency maps, making them unable to explain localized subtle changes [43]. Furthermore, in general, the highlighted or pinpointed attributes need a secondary analysis or interpretation for human-level understanding. For example, when voxels in the subcortical regions of an input MRI are localized as the informative attributes for MCI/AD identification, it is necessary to further analyze whether those regions involve atrophic changes or morphological variations, which can only be done by experts.

Reference-based explanation methods [26], [33], [44], [45] focus on changes in model output with regards to perturbation in input samples. Various perturbation methods that employ strategies such as masking [46], heuristics [33] (*e.g.*, blurring and random noise), using the region of the distractor image as reference for perturbation [27], and synthesized perturbation [26], [44], [45], have been introduced in the literature. One general drawback of these aforementioned attribution-based explanation methods is that they tend to produce similar saliency maps across wrong class labels due to an "attribution vanishing" problem, a phenomenon where the layer-wise explanatory relevance values decrease as the layer levels descend [47].

2.2 Counterfactual Visual Explanations

Recently, more researchers have focused on counterfactual reasoning as a form of visual explanation. Counterfactual explanation refers to analyzing a model's output with regard to hypothetical scenarios. For example, in AD diagnosis, a counterfactual explanation could highlight brain regions that may (hypothetically) cause a normal subject to be diagnosed with a disease when transforming an input image accordingly. VAGAN [31] uses a variant of GAN to synthesize a counterfactual map that transforms an input sample to be classified as another label. However, VAGAN has considerable limitations in its framework. First, for map generation, the true label of an input sample must be known, which is not practically possible in real-world scenarios. Second, VAGAN performs a single-way synthesis only. That is, it generates a counterfactual map that transforms an input originally classified as 'A' to be classified as 'B', but not the reverse. Circumventing the major limitations of VAGAN described above, Bass *et al.* proposed ICAM [32] for producing dual-way counterfactual explanations. However, it cannot be used in tasks with multiple classes of interest and is restricted to being a dual-way explanation.

A possible circumvention to a multi-way explanation would be to combine multiple single-/dual-way models

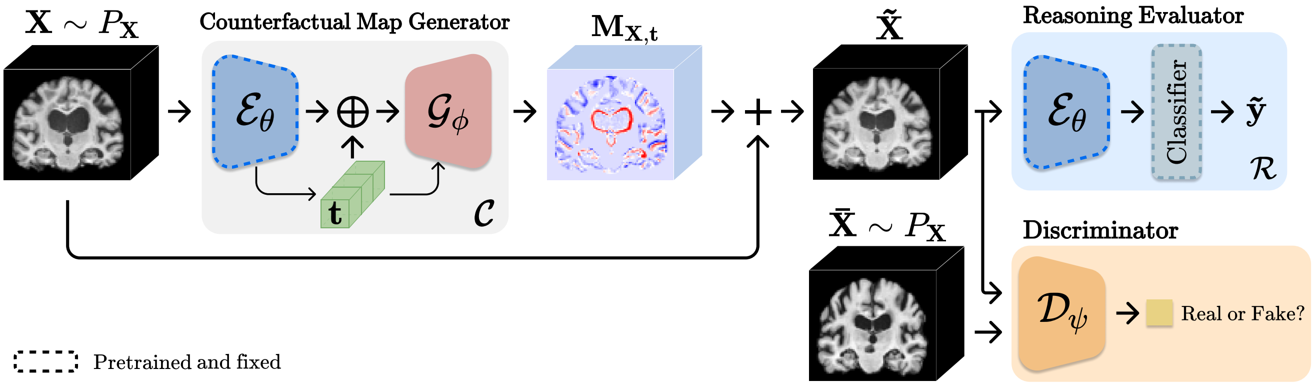

Fig. 2: Schematic overview of the counterfactual map generation to induce the cause of dementia diagnosed from the backbone network. It has major components: counterfactual map generator and reasoning evaluator. The counterfactual map generator synthesizes a counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ conditioned on arbitrary target label $\mathbf{t}$ or posterior probability $\mathbf{t}'$ obtained from the diagnostic model $\mathcal{R}(\mathbf{X})$, while the reasoning evaluator works towards enforcing target label attributes to the synthesized map. Note that $\oplus$ is the operator for channel-wise concatenation and $+$ is the operator for element-wise addition. $\mathbf{X}$ and $\bar{\mathbf{X}}$ are two random instances drawn from the same data distribution, *i.e.*, $\mathbf{X}, \bar{\mathbf{X}} \sim P_{\mathbf{X}}$.

to perform a multi-way explanation. However, one crucial downside of a combination of multiple dual-way implementations compared to a single multi-way implementation is that their outputs do not preserve the associations or relations among the target labels, *e.g.*, clinical stages in the AD spectrum. Note that multiple dual-way implementations generate explainable maps separately and independently without considering the relations among the target classes. Thus, there is no guarantee that the counterfactual maps generated from the multiple independently trained dual-way implementations will preserve the class relations, *i.e.*, CN-MCI-AD in our case. Another limitation of a combination of dual-way models is that it is not fundamentally suitable for real-world scenarios because the true label of an input instance must be known prior to counterfactual map generation. Specifically, given a set of CN vs. MCI, MCI vs. AD, and CN vs. AD dual-way models, it is required to know the true label of an input MRI to select the appropriate dual-way model for counterfactual map generation. Note that given an input MRI of a cognitively normal subject, generating a counterfactual map of MCI or AD with the MCI vs. AD model has no meaning.

It is also noteworthy that VAGAN and ICAM focus on generating images to be classified as another specified target class, rather than elucidating the reasoning or explaining a classifier’s decision. In this work, we propose a novel counterfactual explanation method that can be differentiated from the aforementioned methods as follows: (1) Our proposed method is fundamentally designed to generate counterfactual maps to explain a predictive model’s output in a post-hoc manner. (2) Our proposed method is applicable to a predictive model trained for multi-class classification tasks, *e.g.*, CN *vs.* MCI *vs.* AD and handwritten digit recognition (MNIST), in generating *multi-way* counterfactual maps in a single framework. (3) Our proposed LEAR framework is designed to work with most connectionist models, such as ResNet18 [48], VoxCNN [10], and SonoNet16 [49], for generating counterfactual maps.

2.3 Attention with Guidance

Inspired by the recent successes of deep learning techniques using anatomical landmarks in sMRI, several studies have utilized deep neural networks to guide anatomically and neurologically meaningful regions for brain disease diagnosis [40], [42], [50]. Lian *et al.* [50] proposed a hierarchical fully convolutional network (H-FCN) using anatomical landmarks, which were used as prior knowledge to rule out non-disease-related regions via an attention mechanism, so as to learn discriminative representations more efficiently. The attention-guided HybNet [40] was also proposed to extract discriminative patches and regions from a whole-brain MRI by exploiting CAM extracted from pre-trained models, upon which multi-scale features were jointly trained and fused to construct a hierarchical classification model for AD diagnosis. In the same line of strategies, Li *et al.* [42] proposed an iterative guidance method using CAM for joint pathological region localization and identification for enhancing the diagnostic performance.

While the AD-induced anatomical changes in a brain are subtle, especially in the preclinical or prodromal stages, and are heterogeneous across patients, the aforementioned CAM-based methods can take advantage of only coarse-grained guidance because of the blurry nature of CAM. By contrast, the counterfactual maps obtained from our visual explanation method can provide fine-grained guidance as they represent the minimal source of information to change the clinical label of an input MRI into other ones. By regarding the counterfactual maps as *privileged information*, we devise a novel explanation-guided attention (XGA) module that helps reinforce the generalizability of the predictive network, thus improving its diagnostic performance.

In an effort to categorize the related works of our article, we have categorized some state-of-the-art works into visual explanation, use of attention mechanism, explanation-guided methods, and the ability to reinforce visual explanation, as presented in Table 1. Our work is, to the best of our knowledge, the first that exploits an explanation output to improve the generalization of a diagnostic model reciprocally.

3 Method

In this section, we describe our LEAR framework for visual explanation and reinforcing a diagnostic model. As schematized in Fig. 1, there are two principal units in the framework. The first is an *explanation unit* (EU) that learns a counterfactual map generator $\mathcal{C}$, aimed to visually explain the output of a pre-trained diagnostic model for AD/MCI/CN diagnosis. The other one is a *reinforcement unit* (RU) that, guided by the counterfactual maps generated in EU, updates the learnable parameters of the diagnostic model to improve its generalizability and performance. In addition to these two principal units, our framework also involves a step of pre-training a diagnostic model in a conventional manner, *i.e.*, supervised learning using training samples.

Throughout this article, we denote network models including a diagnostic model, a counterfactual map generation model, and their subnetworks using calligraphic font, while vectors and matrices are denoted by boldface lower and uppercase letters, respectively. The sets are denoted using a typeface style.

Without loss of generality, we assume that a diagnostic model $\mathcal{R}$ is a CNN-based architecture (*e.g.*, ResNet18 [48], VoxCNN [10], and SonoNet16 [49]) and is trained using a whole-brain 3D MRI as input.

3.1 Counterfactual Visual Explanation Model

Given a pre-trained diagnostic model $\mathcal{R}$, we describe our novel visual explanation model $\mathcal{C}$ for the output of the diagnostic model. Formally, the goal of our visual explanation model $\mathcal{C}$ is to infer a counterfactual reasoning map over the output label from a diagnostic model. To this end, we develop a counterfactual map generation method in a GAN framework.

The overall structure for learning our visual explanation model is illustrated in Fig. 2. It has three major modules of a counterfactual map generator (CMG), a reasoning evaluator (RE), and a discriminator (DC). The role of the three modules can be summarized as follows:

- CMG: Given an input MRI sample $\mathbf{X}$ and a target label $\mathbf{t}$, where $\mathbf{t} = [0,1]^{|\mathcal{Y}|}$ is a class probability distribution vector and $|\mathcal{Y}|$ is the size of the class space $\mathcal{Y}$, CMG generates a map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ which, when added to the input $\mathbf{X}$, *i.e.*, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$, causes the transformed image $\tilde{\mathbf{X}}$ to be categorized into the target label $\mathbf{t}$ with high confidence.
- RE: This basically exploits the diagnostic model $\mathcal{R}$ itself. It directly evaluates the effect of the generated counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ in producing the targeted label $\mathbf{t}$, possibly diagnosed differently from the output label of the original input $\mathbf{X}$.
- DC: This helps the CMG to generate an anatomically and morphologically meaningful map, making the transformed image $\tilde{\mathbf{X}}$ realistic.

As RE and DC are, respectively, the network of a diagnostic model and a typical component in GAN, we describe only the CMG in detail.

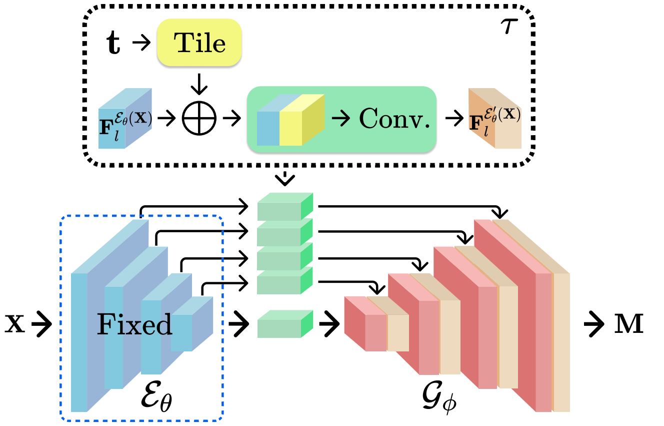
t
Tile
τ
$\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}$
Conv.
$\mathbf{F}_l^{\mathcal{E}'_\theta(\mathbf{X})}$
X
Fixed
$\mathcal{E}_\theta$
$\mathcal{G}_\phi$
M

Fig. 3: Detailed view of the counterfactual map generator (CMG). A target label $\mathbf{t}$ is tiled and channel-wise concatenated to the skip connection. This enables the CMG to condition the counterfactual maps to be conditioned on an arbitrary target condition.

3.1.1 Counterfactual Map Generator (CMG)

The CMG is a variant of Conditional GAN [51] that can synthesize a counterfactual map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ conditioned on a target label $\mathbf{t}$. It consists of an encoder $\mathcal{E}_\theta$ and a generator $\mathcal{G}_\phi$, where the subscripts θ and ϕ denote the tunable parameters of the respective networks. The network design of the encoder $\mathcal{E}_\theta$ and the generator $\mathcal{G}_\phi$ is a variant of U-Net [52] with a tiled target label concatenated to the skip connections, as presented in Fig. 3. Here, we should emphasize that the encoder $\mathcal{E}_\theta$ is taken from the set of layers and the corresponding parameters to extract features in a pre-trained diagnostic model $\mathcal{R}$ with weights θ fixed. Therefore, the encoder $\mathcal{E}_\theta$ is already capable of extracting disease-related features from an input $\mathbf{X}$, thus making our CMG trainable relatively easily and robustly by tuning the parameters of layers other than in the encoder $\mathcal{E}_\theta$ only.

Let $\{\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}\}_{l=1}^{L}$ denote the output feature maps of the L convolution layers in the encoder $\mathcal{E}_\theta(\mathbf{X})$. A given target label $\mathbf{t}$ is concatenated with the feature maps after tiling so that their shapes match the respective feature maps concatenated, *i.e.*, tile $\mathbf{t}$ with the size of $w_l \times h_l \times d_l \times c$ where w_l, h_l, and d_l denote, respectively, the width, height, and depth of a feature map from the l-th convolution block, and c denotes the number of channels. In order to extract better representations of the target label related information, we apply a convolution operation (Conv3D) with a learnable $3 \times 3 \times 3$ kernel, a stride of 1 in each dimension, and zero padding, followed by a nonlinear LReLU activation function as follows (see Fig. 3):

$$\tau(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t}) = \mathrm{LReLU}\left(\mathrm{Conv3D}\left(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})} \oplus \mathrm{Tile}(\mathbf{t})\right)\right) \quad (1)$$

where $\oplus$ denotes an operator of channel-wise concatenation. Then, the target label information included feature maps $\mathbf{F}_l^{\mathcal{E}'_\theta} = \tau(\mathbf{F}_l^{\mathcal{E}_\theta(\mathbf{X})}, \mathbf{t})$ are transmitted to the generator $\mathcal{G}_\phi$ via skip connections. The generator $\mathcal{G}_\phi$ is then able to responsibly synthesize a map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ from the target label informed feature maps as follows:

$$\mathbf{M}_{\mathbf{X},\mathbf{t}} = \mathcal{G}_\phi\left(\mathcal{T}(\mathbf{X}, \mathbf{t})\right) \quad (2)$$

where $\mathcal{T}(\mathbf{X},\mathbf{t}) = \{\tau(\mathbf{F}_1^{\mathcal{E}_\theta(\mathbf{X})},\mathbf{t}),...,\tau(\mathbf{F}_L^{\mathcal{E}_\theta(\mathbf{X})},\mathbf{t})\}$. Finally, we produce a transformed MRI $\tilde{\mathbf{X}}$ by combining the synthesized map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ with an input MRI $\mathbf{X}$ via addition, *i.e.*, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$, which is supposed to be classified as the target label $\mathbf{t}$ by the following RE module, *i.e.*, the diagnostic model $\mathcal{R}$.

Note that in a setting where the target label $\mathbf{t}$ (*e.g.*, CN) is different from the ground-truth label $\mathbf{y}$ (*e.g.*, AD), we may hypothesize that the synthesized map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ visually explains why the input $\mathbf{X}$ was classified to $\mathbf{t}_{\mathbf{X}}$ (*e.g.*, AD), instead of $\mathbf{t}$ (*e.g.*, CN) because $\mathbf{M}_{\mathbf{X},\mathbf{t}}$ highlights the hypothetical regions that contributed to transforming an AD-like MRI $\mathbf{X}$ to a CN-like MRI $\tilde{\mathbf{X}}$.

3.1.2 Counterfactual Visual Explanation Model Training

In this subsection, we define a set of loss functions to train our counterfactual visual explanation model.

Cycle Consistency: In order to encourage the synthesized map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$, which is conditioned on an input $\mathbf{X}$ and a target label $\mathbf{t}$, to be anatomically and morphologically meaningful, we exploit a cycle consistency loss [53] with ℓ_1-norm as follows:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}},\mathbf{t}\sim U(0,|\mathbf{y}|)}\left[\|\mathbf{X}' - \mathbf{X}\|_1\right] \tag{3}$$

where $P_{\mathbf{X}}$ denotes a distribution of MRI samples, $|\mathcal{Y}|$ is the number of classes, $U(\cdot)$ is the one-hot encoded form of a discrete uniform distribution, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$ and $\mathbf{X}' = \tilde{\mathbf{X}} + \mathbf{M}_{\tilde{\mathbf{X}},\mathcal{R}(\mathbf{X})}$. As we propose a way of generating multi-way counterfactual maps, this loss is imperative to synthesize different counterfactual maps for different conditions without suffering from a mode collapse problem [54].

Note that, in following equations, we omit arbitrary target labels $\mathbf{t} \sim U(0,|\mathcal{Y}|)$ from the expectation terms for simplicity.

Adversarial Learning: Inspired by Least Square GAN [55], we adopt the least squares loss function that penalizes samples distant from the discriminator's decision boundary. Using the cycle consistency loss in Eq. (3), the least squares loss needs to be applied to arbitrary real MRI samples $\bar{\mathbf{X}}$, and transformed (*i.e.*, fake) samples $\tilde{\mathbf{X}}$ and $\mathbf{X}'$.

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{\mathcal{D}_\psi} &= \mathbb{E}_{\bar{\mathbf{X}}\sim P_{\mathbf{X}}}\left[(D_\psi(\bar{\mathbf{X}}) - 1)^2\right] \\ &\quad + \frac{1}{2}\left(\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[D_\psi(\tilde{\mathbf{X}})^2 + D_\psi(\mathbf{X}')^2\right]\right)\end{aligned} \tag{4}$$

$$\mathcal{L}_{\text{adv}}^{\mathcal{G}_\phi} = \frac{1}{2}\left(\mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[(D_\psi(\tilde{\mathbf{X}}) - 1)^2 + (D_\psi(\mathbf{X}') - 1)^2\right]\right) \tag{5}$$

This objective function is very suitable for our CMG training because the generated counterfactual maps should neither destroy the input appearance nor ignore the target attribution.

Total Variation: For a more natural synthesis of the counterfactual map generated from CMG and its harmonization with an input sample, we exploit the total variation loss [56] as a regularizer.

$$\begin{aligned}\mathcal{L}_{\text{tv}} = \sum_{i,j,k}\left|\tilde{\mathbf{X}}_{i+1,j,k} - \tilde{\mathbf{X}}_{i,j,k}\right| &+ \left|\tilde{\mathbf{X}}_{i,j+1,k} - \tilde{\mathbf{X}}_{i,j,k}\right| \\ &+ \left|\tilde{\mathbf{X}}_{i,j,k+1} - \tilde{\mathbf{X}}_{i,j,k}\right|\end{aligned} \tag{6}$$

where $\tilde{\mathbf{X}} = \mathbf{X}+\mathbf{M}_{\mathbf{X},\mathbf{t}}$, and i, j, and k are indices of each axis in the 3D coordinate of a volumetric image, respectively.

Sparsity in a Counterfactual Map: From the interpretability and identity preservation standpoints, it is crucial to regularize the dense counterfactual map to highlight only the essential regions necessary for counterfactual reasoning. To this end, we also impose an elastic regularization on the synthesized counterfactual map as follows:

$$\mathcal{L}_{\text{map}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\lambda_1\|\mathbf{M}_{\mathbf{X},\mathbf{t}}\|_1 + \lambda_2\|\mathbf{M}_{\mathbf{X},\mathbf{t}}\|_2\right] \tag{7}$$

where λ_1 and λ_2 are the weighting hyperparameters.

Correctness of Counterfactual Reasoning: To ensure that the transformed image $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{M}_{\mathbf{X},\mathbf{t}}$ is correctly conditioned on the target label $\mathbf{t}$, we include a classification loss function as follows:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\text{CE}\left(\mathbf{t},\tilde{\mathbf{y}}\right))\right] \tag{8}$$

where CE denotes a cross-entropy function, and $\tilde{\mathbf{y}} = \mathcal{R}(\tilde{\mathbf{X}})$ is a softmax activated class probability distribution vector.

Conceptually, the role of the diagnostic model $\mathcal{R}$ is similar to that of a discriminator $\mathcal{D}_\psi$, but their objective is very different. While a discriminator $\mathcal{D}_\psi$ learns to distinguish between real and fake samples, the diagnostic model $\mathcal{R}$ is already trained to classify the input samples correctly. Thus, the diagnostic model $\mathcal{R}$ provides a deterministic guidance for the generator to produce a target-directed counterfactual map, while the discriminator $\mathcal{D}_\psi$ plays a min-max game with a generator $\mathcal{G}_\phi$ in an effort to produce more realistic samples.

3.1.3 Total Loss Function

We define the total loss function for counterfactual map generation as follows:

$$\mathcal{L}_{\text{CMG}} = \lambda_3\mathcal{L}_{\text{adv}}^{\mathcal{G}_\phi} + \lambda_4\mathcal{L}_{\text{adv}}^{\mathcal{D}_\psi} + \lambda_5\mathcal{L}_{\text{cyc}} + \lambda_6\mathcal{L}_{\text{cls}} + \lambda_7\mathcal{L}_{\text{tv}} + \mathcal{L}_{\text{map}} \tag{9}$$

where λ_* values are the hyperparameters of the model ($\lambda_{1,2}$ in Eq. (7)). We empirically tuned λ such that the magnitude of gradients of each loss term is roughly balanced (see Supplementary S1.3). An ablation study of these loss functions is provided in Section 4.1.2. Overall, each of the loss terms improves the quality of generated counterfactual maps.

It should be noted that during training, we share and fix the weights of the encoder $\mathcal{E}_\theta$ of the CMG with the RE module's feature extractor $\mathcal{E}_\theta$ to ensure that the attribution is consistent throughout the generative process.

3.2 Reinforcement Representation Learning

In this article, we hypothesize that the set of counterfactual maps synthesized by our CMG along with a diagnostic model can be a vital source of knowledge of anatomical or morphological changes relevant to AD, inferred in a data-driven manner. Such data-driven knowledge is comparable to the conventional neuroscientific knowledge mostly acquired from a group statistical analysis in a univariate manner [3]. Note that the diagnostic model is trained with the aim of classifying samples having different clinical labels, *e.g.*, CN, MCI, AD, by discovering generalizable and discriminative patterns inherent in samples. Our proposed CMG is designed and trained to detect such generalizable and discriminative patterns in an input sample to explain the diagnostic model's output via counterfactual reasoning.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3197845

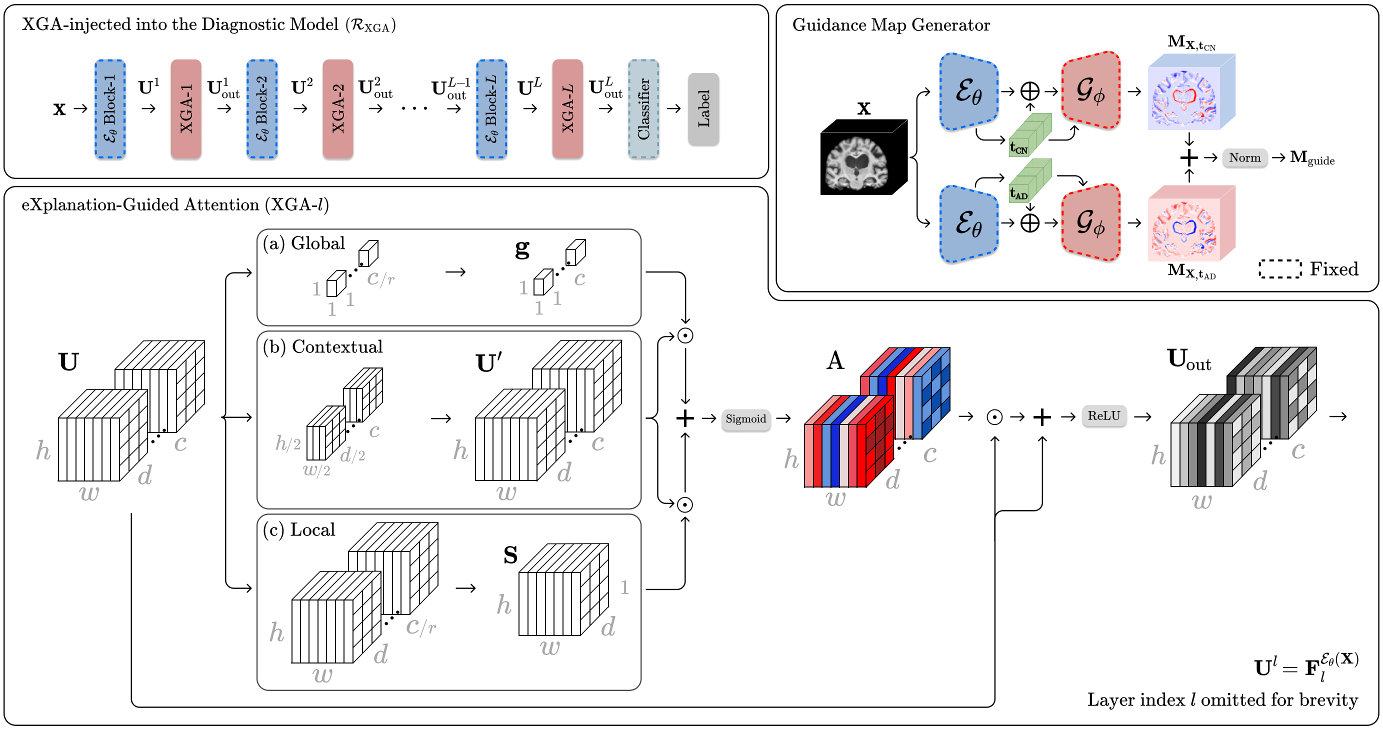
XGA-injected into the Diagnostic Model ($\mathcal{R}_{\text{XGA}}$)
$\mathbf{x}$
$\mathcal{E}_\theta$ Block-1
$\mathbf{U}^1$
XGA-1
$\mathbf{U}^1_{\text{out}}$
$\mathcal{E}_\theta$ Block-2
$\mathbf{U}^2$
XGA-2
$\mathbf{U}^2_{\text{out}}$
$\mathbf{U}^{L-1}_{\text{out}}$
$\mathcal{E}_\theta$ Block-L
$\mathbf{U}^L$
XGA-L
$\mathbf{U}^L_{\text{out}}$
Classifier
Label
Guidance Map Generator
$\mathbf{x}$
$\mathcal{E}_\theta$
$\mathcal{G}_\phi$
$\mathbf{t}_{\text{CN}}$
$\mathbf{t}_{\text{AD}}$
$\mathbf{M}_{\mathbf{x},\mathbf{t}_{\text{CN}}}$
$\mathbf{M}_{\mathbf{x},\mathbf{t}_{\text{AD}}}$
Norm
$\mathbf{M}_{\text{guide}}$
Fixed
eXplanation-Guided Attention (XGA-l)
$\mathbf{U}$
(a) Global
$\mathbf{g}$
c/r
c
(b) Contextual
$\mathbf{U}'$
$h/2$
$w/2$
$d/2$
(c) Local
$\mathbf{S}$
h
w
d
c
c/r
1
Sigmoid
$\mathbf{A}$
ReLU
$\mathbf{U}_{\text{out}}$
$\mathbf{U}^l = \mathbf{F}_l^{\mathcal{E}_\theta}(\mathbf{X})$
Layer index l omitted for brevity

Fig. 4: Schematic overview of the explanation-guided attention (XGA) module with a guidance map. A guidance map is a supervision for the XGA module that assists in focusing on regions of pathological and morphological changes caused by dementia on the whole-brain. XGA module learns and integrates locally subtle changes and globally discriminative structural changes that can optionally be supervised by the guidance map. Note that $\odot$ is the operator for element-wise multiplication and $+$ is the operator for element-wise addition.

3.2.1 Guidance Map Generation

Based on these considerations, we propose to exploit the counterfactual maps as guidance to reinforce the diagnostic model's representations. Specifically, we generate the counterfactual maps of an input sample with the target labels of $\mathbf{t}_{\text{CN}}$ and $\mathbf{t}_{\text{AD}}$, *i.e.*, the most normal and the most AD-like brains with regard to the input sample. For example, in a 3-class classification task of CN *vs.* MCI *vs.* AD, $\mathbf{t}_{\text{CN}} = [1, 0, 0]$ is the class probability distribution vector of CN and $\mathbf{t}_{\text{AD}} = [0, 0, 1]$ is the class probability distribution vector of AD. Assuming that these two counterfactual maps jointly represent the localized AD-sensitive regions over the whole brain, we build a guidance map $\mathbf{M}_{\text{guide}}$ by combining them as follows:

$$\mathbf{M}_{\text{guide}} = \text{MinMax}\left(|\mathbf{M}_{\mathbf{X},\mathbf{t}_{\text{CN}}}| + |\mathbf{M}_{\mathbf{X},\mathbf{t}_{\text{AD}}}|\right) \tag{10}$$

where $|\cdot|$ is an absolute operation and $\text{MinMax}(\cdot)$ denotes a min-max normalization in a voxel-wise manner. Thus, the absolute term in the guidance map $\mathbf{M}_{\text{guide}}$ allows the use of attentive values in both the extreme cases of most normal brain and most AD-like brain (because negative values of $\mathbf{M}_{\mathbf{X},\mathbf{t}_{\text{CN}}}$ highlight the most normal regions of the brain compared to AD-affected brains, while positive values of $\mathbf{M}_{\mathbf{X},\mathbf{t}_{\text{AD}}}$ highlight the most AD-like regions of the brain). This guidance map is then used to reinforce the representational power of the layers' outputs in the diagnostic model by modulating them via the attention mechanism described below. Note that we do not include $\mathbf{M}_{\mathbf{X},\mathbf{t}_{\text{MCI}}}$ in Eq. (10) because of its redundancy in creating a guidance map as MCI is an intermediate stage between CN and AD in the AD spectrum.

3.2.2 Explanation-Guided Attention

In order to exploit the explanation-induced knowledge of the anatomical and morphological changes for AD diagnosis, we devise an explanation-guided attention (XGA) module by regarding the counterfactual maps as *model-driven privileged information* during training. Specifically, we inject a self-attention module that adaptively modulates the layer outputs in the diagnostic model (Fig. 4).

Let $\mathbf{U}^l$ be an output feature map of the l-th layer in the diagnostic model $\mathcal{R}$, *i.e.*, $\mathbf{U}^l = \mathbf{F}_l^{\mathcal{E}_\theta}$, and $\mathbf{A}^l$ its resulting attention map, whose computation is detailed below. Note, it is expected that the attention map $\mathbf{A}^l$ produces the higher attentive values, where the higher explanation values are in the guidance map $\mathbf{M}_{\text{guide}}$, obtained by Eq. (10), for an input sample. Thereby, the AD-sensitive regions, guided by the counterfactual maps, are excited with the discriminative representations while other regions are inhibited, thus reinforcing the feature representations in the diagnostic model.

We base our computation to estimate the attention map on the global-and-local (GALA) module [57], which consists of global and local operators, as presented in Fig. 4 (a), (c), respectively. Specifically, our XGA module adapts the local and global attention operators of GALA and improves them using a contextual attention operator. Using this simple modification to the GALA attention mechanism, our XGA module achieved about 11% accuracy improvement in 3-class diagnosis experiments (refer to Supplementary S7).

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3197845

Basically, the following operations can be applied to different layers equally. Hereafter, we omit the superscript l of a layer index to reduce clutter. Our XGA modulates an input feature map $\mathbf{U}$ with an attention map $\mathbf{A}$ of the same dimension as $\mathbf{U}$. That is, $\mathbf{U}, \mathbf{A} \in \mathbb{R}^{w \times h \times d \times c}$, where w, h, d, and c are the spatial width, height, depth, and number of feature channels, respectively.

Global Attention: First, we account for the global attention in the XGA module by exploiting the squeeze-and-excitation technique [57], [58]. To obtain the global feature attention vector $\mathbf{g} \in \mathbb{R}^{1 \times 1 \times 1 \times C}$, we first obtain a channel descriptor $\mathbf{d} = [d_c]_{c=1}^{C}$ by calculating the summary statistics of the c-th channel via global average pooling, *i.e.*, $d_c = \frac{1}{WHD} \sum_{w=1}^{W} \sum_{h=1}^{H} \sum_{d=1}^{D} U_{whdc}$. As the channel descriptor $\mathbf{d}$ includes information obtained from the full receptive field, it can be thought of as carrying the importance of the respective channel with respect to the global information. This is followed by a two-layer neural network that non-linearly transforms the channel descriptor to explicitly model the inter-dependencies among the channels as follows:

$$\mathbf{g} = \mathbf{W}_{\text{expand}}(\text{ReLU}(\mathbf{W}_{\text{c-shrink}}(\mathbf{d}))). \tag{11}$$

where $\mathbf{W}_{\text{c-shrink}} \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_{\text{expand}} \in \mathbb{R}^{C \times \frac{C}{r}}$ are the shrinking and expansion operations, respectively, and r is a ratio hyperparameter.

Local Attention: Second, we consider a local saliency component to compute the local feature attention $\mathbf{S}$. Unlike the global attention, the local feature attention map $\mathbf{S}$ focuses on "where" a crucial part locates, complementing the global attention. While retaining the spatial dimensions, we conduct two consecutive convolution operations along the channel dimension with a non-linear transformation in-between to enhance the complexity as follows:

$$\mathbf{S} = \mathbf{W}_{\text{collapse}} * (\text{ReLU}(\mathbf{W}_{\text{d-shrink}} * \mathbf{U})) \tag{12}$$

where $*$ denotes convolution, $\mathbf{W}_{\text{d-shrink}} \in \mathbb{R}^{1 \times 1 \times 1 \times C \times \frac{c}{r}}$ and $\mathbf{W}_{\text{collapse}} \in \mathbb{R}^{1 \times 1 \times 1 \times \frac{c}{r} \times 1}$ are learnable parameters. This local attention is used to generate an inter-feature attention map by allowing for the channel-wise relationship of features.

Contextual Attention: Along with the global and local attention operators of GALA described above, our XGA also involves a contextual attention operator. It is designed to utilize the contextual information from a larger receptive field. To this end, we first conduct a dilated convolution [59], which has the effect of taking into account features of enlarged field of view and reducing the map size, followed by a non-linear transformation. We then up-scale its output back to the input size of $\mathbf{U}$ as follows:

$$\mathbf{U}' = \text{Up}(\text{ReLU}(\mathbf{W}_{\text{reduction}} *_d \mathbf{U})) \tag{13}$$

where $*_d$ denotes dilated-convolution and $\text{Up}(\cdot)$ is an operator for trilinear up-scaling back to the original spatial dimensions of $\mathbf{U}$.

Finally, the global, local, and contextual module outputs are integrated to produce the attention mask $\mathbf{A} \in \mathbb{R}^{w \times h \times d \times c}$ after tiling $\mathbf{g}$ and $\mathbf{S}$ to form $\mathbf{G}^*$, $\mathbf{S}^* \in \mathbb{R}^{w \times h \times d \times c}$ owing to their differences in size, as follows:

$$\mathbf{A} = \sigma\left(\mathbf{G}^* \odot \mathbf{U}' + \mathbf{S}^* \odot \mathbf{U}'\right) \tag{14}$$

where σ denotes a sigmoid activation function, $\odot$ denotes element-wise multiplication, and $+$ denotes element-wise addition. The attention mask $\mathbf{A}$ plays the role of excitation and inhibition of the input feature map $\mathbf{U}$ with a skip connection as follows:

$$\mathbf{U}_{\text{out}} = \text{ReLU}(\mathbf{U} + (\mathbf{U} \odot \mathbf{A})). \tag{15}$$

3.2.3 XGA Learning

Inspired by Linsley *et al.* [57], we define the loss using the cross-entropy function regularized by the attention-guidance penalty Ω_{XGA} to train the parameters $\mathcal{R}_\omega = \{\mathbf{W}^l_{\text{c-shrink}}, \mathbf{W}^l_{\text{expand}}, \mathbf{W}^l_{\text{collapse}}, \mathbf{W}^l_{\text{d-shrink}}, \mathbf{W}^l_{\text{reduction}}\}_{l=1}^{L}$ for XGA modules injected next to the every convolution layer of the diagnostic model as follows:

$$\mathcal{L}_{\text{XGA}} = \mathbb{E}_{\mathbf{X} \sim P_{\mathbf{X}}}[\text{CE}(\mathbf{y}, \mathcal{R}_{\text{XGA}}(\mathbf{X}))] + \lambda_8 \Omega_{\text{XGA}} \tag{16}$$

$$\Omega_{\text{XGA}} = \sum_{l \in L} \left\| \frac{\bar{\mathbf{M}}^l_{\text{guide}}}{\|\bar{\mathbf{M}}^l_{\text{guide}}\|_2} - \frac{\bar{\mathbf{A}}^l(\mathbf{X})}{\left\|\bar{\mathbf{A}}^l(\mathbf{X})\right\|_2} \right\|_2 \tag{17}$$

where Ω_{XGA} is a scalar value for attention-guidance penalty, $\mathcal{R}_{\text{XGA}}$ denotes the diagnostic model $\mathcal{R}$ with XGA modules injected, $\bar{\mathbf{A}}^l(\mathbf{X}) \in \mathbb{R}^{w \times h \times d \times 1}$ is the compression of $\mathbf{A}^l(\mathbf{X}) \in \mathbb{R}^{w \times h \times d \times c}$ with channel-wise ℓ_2-norm values, $\bar{\mathbf{M}}^l_{\text{guide}}$ is a trilinear-interpolated form of $\mathbf{M}^l_{\text{guide}}$ to be the same size of $\bar{\mathbf{A}}^l(\mathbf{X})$, and λ_8 is a weighting hyperparameter. While training the parameters of the XGA modules, we fix the other model parameters of the diagnostic model. With regards to the attention-guidance penalty Ω_{XGA}, as described above, we expect that the attention map $\mathbf{A}^l(\mathbf{X})$ of an input sample $\mathbf{X}$ outputs higher values for excitation, where the higher explanation values are in the guidance map $\mathbf{M}_{\text{guide}}$, and lower values for inhibition otherwise. It is noteworthy that even though we regularize the model training by applying the same guidance map for attention over layers, the attention maps of different layers still help find rich and diverse features because of the residual operation in Eq. (15) and the difference in resolution over layers. Thus, the XGA module helps emphasize the features in the attended regions while still considering features in non-attended regions for better layer-wise representations.

3.3 Iterative Explanation-Reinforcement Learning

Finally, we introduce an iterative explanation-reinforcement learning scheme that enhances the quality of visual explanation as well as the performance of the diagnostic model as follows:

Phase 1: CMG training

$$\min_{\mathcal{G}_\phi, \mathcal{D}_\psi} \mathcal{L}_{\text{CMG}}, \tag{18}$$

Phase 2: XGA training

$$\min_{\mathcal{R}_\omega} \mathcal{L}_{\text{XGA}}. \tag{19}$$

In this iterative training scheme, Phase 1 and Phase 2 are repeated sequentially. During the first iteration of the optimization, we use the original definition of counterfactual reasoning map $\mathbf{M}_{\mathbf{X},\mathbf{t}}$, *i.e.*, Equation (2), as XGA has not

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3197845

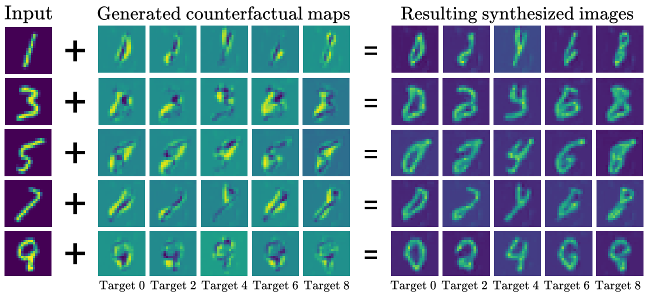

Fig. 5: Examples of counterfactual maps for the MNIST dataset. The resulting synthesized image is an addition between an input and its corresponding counterfactual map (blue and yellow denote, respectively, subtraction and addition of the respective pixel values, *i.e.*, deletion and addition of areas to be a target-labeled digit) conditioned on a target label.

reinforced the diagnostic model yet. For second and later iterations, we redefine the counterfactual map as follows:

$$\mathbf{M}_{\mathbf{X},\mathbf{t}} := \mathcal{G}_\phi\left(\mathcal{T}'(\mathbf{X},\mathbf{t})\right) \tag{20}$$

where $\mathcal{T}'(\mathbf{X},\mathbf{t}) = \{\tau(\mathbf{F}_1^{\mathcal{E}_{\theta,\omega}(\mathbf{X})},\mathbf{t}),...,\tau(\mathbf{F}_L^{\mathcal{E}_{\theta,\omega}(\mathbf{X})},\mathbf{t})\}$, and $\{\mathbf{F}_l^{\mathcal{E}_{\theta,\omega}(\mathbf{X})}\}_{l=1}^{L}$ denote the output feature maps of the L convolution layers in the encoder $\mathcal{E}_{\theta,\omega}(\mathbf{X})$ from the XGA-injected diagnostic model $\mathcal{R}_{\text{XGA}}$. Note that parameters of the pre-trained diagnostic model, *i.e.*, θ, is fixed during all phases and all iterations.

4 Experimental Settings and Results

In this section, we (1) analyze and validate the visual explanation results of our counterfactual reasoning map; (2) show the effectiveness of our LEAR framework in reinforcing the diagnostic models; and (3) apply our LEAR framework to baseline and state-of-the-art diagnostic models to demonstrate its portability.

4.1 Counterfactual Reasoning

4.1.1 Toy Example: MNIST Classifier

In order to help the readers' understanding of a visual explanation method using counterfactual reasoning maps, we present the visual explanation of an MNIST classifier, owing to its intuitiveness.

Dataset and Implementation

MNIST [60] is a gray-scale handwritten digit image dataset that, we believe, is suitable for the proof-of-concept of various visual explanation methods. For the preparation of the dataset, we utilized the data split provided by the dataset publisher [60] and applied min-max normalization. For the classifier model, we re-implemented and pre-trained the model proposed by Kim *et al.* [61] with minor modifications (*e.g.*, kernel and stride size) to accommodate the smaller image size of the MNIST dataset. More details on the implementation are in Supplementary S1.1.

Results and Analysis

Fig. 5 shows examples of the generated counterfactual (CF) maps and the resulting synthesized images towards five targeted classes (*i.e.*, 0, 2, 4, 6, 8) from six different input images. Note that our CMG successfully produced counterfactual visual explanations, indicating which pixels should be deleted (blue) or added (yellow) to be the different target classes, in *multiple* hypothetical scenarios.

We emphasize the importance of visual explanation in *hypothetical scenarios* as it can provide users with an intuitive understanding on *"what if* $\mathbf{X}$ *was* $\mathbf{X}^*$*?"* In this sense, a CF map should transform an input sample $\mathbf{X}$ to be dependent only on the targeted hypothetical scenario and independent to any other artifacts. Our experiment on the MNIST dataset demonstrates this ability to isolate attribution to only the targeted label because the transformed image maintains the style of the input image while successfully being transformed to a target digit. For example, for transforming an image of the digit "3" to a target digit "8", we can observe that the contours of the original "3" image are maintained while new contours are added to form a digit "8". Likewise, for transforming an image of the digit "0" to a target digit "4", we can observe that the upper and bottom arcs were removed to form a digit "4" with the rest of the arcs maintained. This ability to isolate targeted conditions allows our CMG module to be reliably applied to a medical task in the next subsection.

4.1.2 Alzheimer's Disease Classifier

Dataset and Implementation

The ADNI dataset, collated by the Alzheimer's Disease Neuroimaging Initiative [62], is used for the following experiments. The ADNI dataset is highly challenging as it is practice-oriented in real-world medical applications and its images feature subtle and diverse morphological changes. It consists of 3D structural magnetic resonance imaging (sMRI) of various subject groups ranging from cognitive normal (CN) to mild cognitive impairment (MCI) to Alzheimer's disease (AD). We further split MCI subjects into two sub-groups of progressive MCI (pMCI) for MCI subjects who have converted to AD within 36 months of screening and stable MCI (sMCI) for those who remained in the MCI group within 36 months of screening. Specifically, we have utilized 431 CN subjects, 497 sMCI subjects, 251 pMCI subjects, and 359 AD subjects in ADNI-1 and ADNI-2 studies. Some subjects had multiple MRIs acquired during the span of their life, but we have only selected their baseline MRIs. Thus, 1,538 images are used in our experiments. For three-class experiments, sMCI and pMCI subjects were considered as MCI subjects. Note that, in our experiments, prodromal stages from CN to AD are sMCI and pMCI, with the latter generally considered more severe.

We used a five-fold cross-validation setting for all experiments, and used the same indices for all the comparison methods. We made sure there was no data leakage while training the backbone diagnostic models, CMG optimization, XGA optimization, and iterative optimization.

ResNet18 [48] baseline and various state-of-the-art diagnostic models were re-implemented for the encoder $\mathcal{E}_\theta$ to demonstrate the generalizability of our LEAR framework. Note that, unless specified otherwise, we utilize the three-class ResNet18 model as the backbone diagnostic model for the experiments in this section. The decoder $\mathcal{G}_\phi$ in the CMG

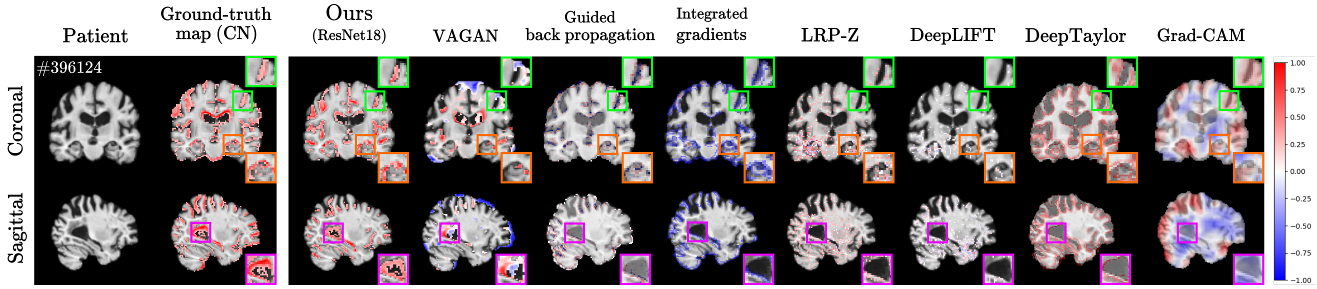

Fig. 6: Example of counterfactual maps for the ADNI dataset (Subject ID 024_S_0985, Image ID on top left corner). Purple, green, and orange boxes visualize ventricular, cortex, and hippocampus regions, respectively.

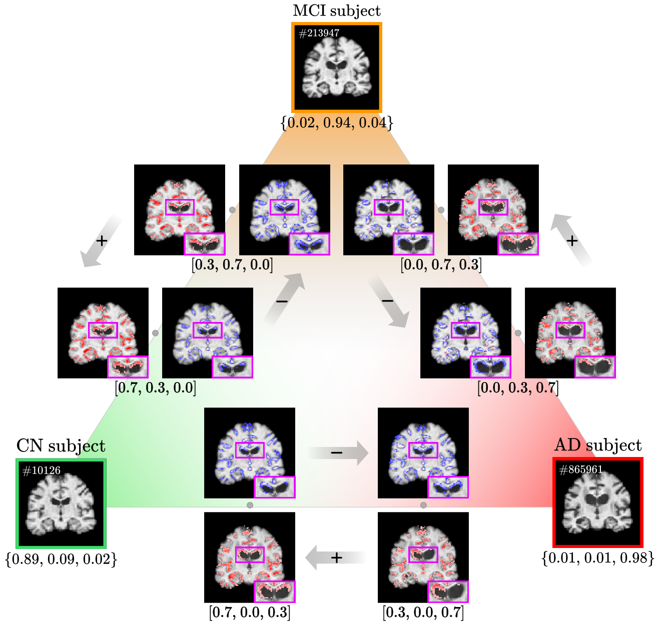

Fig. 7: Example of counterfactual map conditioned on interpolated target labels (Subject ID 123_S_0106, Image ID on top left corner). The purple boxes correspond to the ventricular region. Parentheses $\{\cdot\}$ and $[\cdot]$ for condition indicate the posterior probability and a target condition, respectively. The +/- signs above the gray arrows denote, respectively, NCC(+) and NCC(-). Refer to Supplementary Fig. S1 for a more detailed interpolation result of disease progression.

has the same network design as the encoder $\mathcal{E}_\theta$ with pooling layers replaced by up-sampling layers. We have also utilized the structure of encoder $\mathcal{E}_\theta$ as the DC module $\mathcal{D}_\psi$ identically in all experiments. More details on the implementation, ADNI dataset, and sMRI preprocessing are provided in Supplementary S1.2 and S2.1.

Results and Analysis

In order for qualitative and quantitative evaluation with regard to the visual explanation, we used the longitudinal samples of 12 subjects in ADNI-1/-2, from which the ground-truth maps were created to indicate morphological changes in sMRI according to changes in clinical diagnosis. Details about the longitudinal samples and creating ground-truth maps are given in Supplementary S2.2. It should be noted that none of these images shown there were used in any of our model training procedures.

(*AD→CN Counterfactual Maps*) For visual explanation of a trained three-class diagnostic model, we applied our CMG and other comparative methods in the literature. Fig. 6 illustrates their respective results to explain why the input image was diagnosed into AD, instead of CN. Notably, our proposed CMG showed the best matching result to the ground-truth map by detecting and highlighting the ventricle enlargement and cortical atrophies. These visual explanations are consistent with the existing clinical neuroscience studies [63], [64], [65].

The CF map generated by LRP-Z [23] and DeepLIFT [25] does not clearly show the class discriminative regions. We observe that these approaches focus only on the left hippocampus area while ignoring the right hippocampus area (orange box). Unlike other gradient-based approaches, Guided backpropagation [21], Integrated gradients [22], and DeepTaylor [24] methods showed some traces of counterfactual reasoning across the image, but unnecessary attributions were observed at the edge or morphological boundaries of the brain. Even though Grad-CAM [21] has shown the class-discriminative visualization, this result slightly captures the coarse regions.

GAN-based models (*e.g.*, ours and VAGAN [31]) achieve superior results in comparison to other visual explanation methods. However, VAGAN is only successful in mimicking the hypertrophy in the hippocampus regions while failing to capture the increased cortical thickness (in fact, it decreased the cortical thickness in blue-colored regions). In contrast, our method captures almost every subtle region where the cortical thickness was increased while successfully capturing the reduced ventricular and the hypertrophy in the hippocampus. Thus, our CMG module is able to visually explain class-discriminative and fine-grained regions of the brain.

(*CN↔MCI↔AD Counterfactual Maps*) In addition to our CMG's ability to produce high-quality CF maps, it is also capable of generating counterfactual explanation maps with regard to diverse conditions in the AD spectrum, which, to the best of our knowledge, cannot be done by the existing comparable methods. We generated *multi-way* CF maps with interpolation-based target conditions setting among the three classes of CN, MCI, and AD and illustrated the results in Fig. 7. In this figure, using longitudinal samples of a subject (Subject ID 123_S_0106) who had experienced all

TABLE 2: Normalized Cross-Correlation (NCC) scores with comparison methods on the ADNI dataset. We differentiated NCC scores for each generation direction of the counterfactual map. The +/- signs indicate different directions of the counterfactual map (see Fig. 7).

Scenario	**CN ↔ MCI**		**MCI ↔ AD**		**CN ↔ AD**	
	NCC(+)	**NCC(-)**	**NCC(+)**	**NCC(-)**	**NCC(+)**	**NCC(-)**
LRP-Z [23]	0.005	0.005	0.006	0.004	0.008	0.005
Integrated Gradients [22]	0.006	0.007	0.007	0.007	0.006	0.005
DeepLIFT [25]	0.004	0.005	0.006	0.004	0.005	0.004
Guided Backprop [21]	0.199	0.158	0.212	0.163	0.239	0.204
DeepTaylor [24]	0.143	0.172	0.112	0.108	0.132	0.118
Grad-CAM [21]	0.201	0.188	0.215	0.227	0.227	0.214
VA-GAN [31]	0.283	0.186	0.285	0.257	0.317	0.298
Ours	**0.364**	**0.289**	**0.299**	**0.297**	**0.366**	**0.312**

the clinical stages of CN, MCI, and AD over several years, we produced CF maps under various target conditions. For example, a target condition of $\tilde{\mathbf{t}} = [0.3, 0.7, 0]$, where each element accounts for the probability of belonging to the CN, MCI, and AD group, respectively, was used to transform an MCI image $\mathbf{X}_{\text{MCI}}$ to a prodromal CN-like image using the CF map $\mathbf{M}_{\mathbf{X}_{\text{MCI}},\tilde{\mathbf{t}}}$ (first image in the top row of Fig. 7), and the same target condition could be used to transform a CN image $\mathbf{X}_{\text{CN}}$ to a prodromal MCI-like image using the CF map $\mathbf{M}_{\mathbf{X}_{\text{CN}},\tilde{\mathbf{t}}}$ (second image in the top row).

Although linearly interpolating between stages of a disease is not a pathologically sound procedure for analyzing the intermediate stages of a disease, our CMG produces sub-optimal[‡] CF maps in that biomarkers gradually increase or decrease with regards to a given target condition (results on interpolation in finer steps are in Supplementary Fig. S1). For example, the size of the ventricle (purple boxes) gradually reduces when interpolating from $\mathbf{X}_{\text{MCI}}$ to $\mathbf{X}_{\text{CN}}$ (*i.e.*, the "+" direction), and gradually enlarges when interpolating in the opposite direction (*i.e.*, the "-" direction). Additionally, we can see that the magnitude of the CF maps $\mathbf{M}_{\mathbf{X}_{\text{CN}},\mathbf{t}_{\text{MCI}}}$ (for transforming $\mathbf{X}_{\text{CN}}$ to $\mathbf{X}_{\text{MCI}}$) and $\mathbf{M}_{\mathbf{X}_{\text{MCI}},\mathbf{t}_{\text{AD}}}$ approximately add up to the CF map $\mathbf{M}_{\mathbf{X}_{\text{CN}},\mathbf{t}_{\text{AD}}}$ (*i.e.*, $\mathbf{M}_{\mathbf{X}_{\text{CN}},\mathbf{t}_{\text{AD}}} \approx \mathbf{M}_{\mathbf{X}_{\text{CN}},\mathbf{t}_{\text{MCI}}} + \mathbf{M}_{\mathbf{X}_{\text{MCI}},\mathbf{t}_{\text{AD}}}$). This indicates that our CMG is able to successfully capture the two extreme tails of most normal brain and most AD-like brain, which strengthens our motivation for using these maps as the source of attention in the XGA module (see Section 3.2.1).

(Quantitative Evaluation) To quantitatively assess the quality of our generated CF maps, we calculated the normalized cross-correlation (NCC) score between generated CF maps and ground-truth maps by following [31]. The NCC score measures the similarity between two samples in a normalized setting where higher NCC scores denote higher similarity. Thus, NCC can be helpful when two samples have a different magnitude of signals. Here, we denote the ground-truth maps and CF maps for transforming CN←MCI, MCI←AD, CN←AD as the "+" direction and CN→MCI, MCI→AD, CN→AD as the "-" direction (see Fig. 7), and calculate NCC(+) and NCC(-) for each.

TABLE 3: Normalized Cross-Correlation (NCC) scores in an ablation study of the loss terms in Eq. (9).

Components				**CN ↔ MCI**		**MCI ↔ AD**		**CN ↔ AD**	
$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{adv}}^{\mathcal{D}_\psi,\mathcal{G}_\phi}$	$\mathcal{L}_{\text{map}}$	$\mathcal{L}_{\text{tv}}$	**NCC(+)**	**NCC(-)**	**NCC(+)**	**NCC(-)**	**NCC(+)**	**NCC(-)**
	✓			0.048	0.055	0.045	0.063	0.066	0.081
✓				0.219	0.199	0.196	0.203	0.231	0.228
✓	✓			0.328	0.256	0.254	0.247	0.331	0.289
✓		✓	✓	0.269	0.237	0.217	0.208	0.278	0.241
✓	✓		✓	0.360	0.281	0.289	0.291	0.348	0.292
✓	✓	✓		0.358	0.279	0.285	0.287	0.353	0.305
✓	✓	✓	✓	**0.364**	**0.289**	**0.299**	**0.297**	**0.366**	**0.312**

The scores for LRP-Z [23] and DeepLIFT [25] are understandably low because they can only capture the least number of class-discriminative features as seen in Fig. 6. Integrated Gradients [21] can capture the class-discriminative features in a group-wise manner, *i.e.*, the values of their CF maps do not differ significantly for different subjects, and so their NCC score, which is a subject-wise correlation score, is very low. Guided Backprop [21], DeepTaylor [24], and Grad-CAM [21] can capture some class-discriminative features, but only in a coarse-grained manner. We found that VA-GAN [31] has captured some meaningful regions for disease localization. However, NCC scores of our proposed CMG are higher than VAGAN (Table 2) because our CF maps can localize biomarkers throughout the brain, while the CF maps of VAGAN fail to capture the class-discriminative features in the cortical regions. Unlike the competing methods, which are built on top of binary classifiers, our LEAR framework can fully utilize various backbone diagnostic models (*e.g.*, ResNet18, VoxCNN, and SonoNet16) for binary and multi-class classification tasks. The full and comprehensive results are provided in Supplementary S5.1.

One interesting phenomenon across methods is the lower NCC scores in the "-" direction, *i.e.*, NCC(-). A simple hypothesis we made was that more (difficult) processes are required for subtracting, which happens mostly in the "-" direction, than for adding certain regions of a brain. For example, a baseline CN image tends to have more gray matter (*i.e.*, gray colored tissues) in certain biomarker regions than its progressed AD image. These gray matter regions also tend to contain more complex morphological features than other (*i.e.*, white matter and cerebrospinal fluid) regions, which makes transforming to gray matter (*i.e.*, a "-" operation) more difficult given that morphological features need to be drawn on top of those regions. With that said, our LEAR framework could mitigate this gap between NCC(+) and NCC(-) using the guidance map $\mathbf{M}_{\text{guide}}$ which allows our framework to account for the both extreme cases of most normal and most AD-like brains.

(Ablation Study) As our ablation study show in Table 3, each loss term has different roles in generating counterfactual maps. Most notably, ablating the classifier loss $\mathcal{L}_{\text{cls}}$ results in a considerable drop in NCC scores, indicating that it is one of the most crucial components of LEAR in conditioning the counterfactual maps with regards to the target label $\mathbf{t}$. When missing the total variation loss $\mathcal{L}_{\text{tv}}$ in Eq. (6), the sparsity loss $\mathcal{L}_{\text{map}}$ in Eq. (7), and the GAN losses in Eq. (4) and Eq. (5), it causes performance degradation overall but with different amounts. In particular, $\mathcal{L}_{\text{tv}}$ is responsible

[‡] It is sub-optimal because the progression of disease is not a linear process, but our target condition is.

TABLE 4: Comparison of performance (ACC) among the backbone, augmentation, and the attention with guidance on ADNI dataset.

Setting	**ResNet18**		
	backbone	**augmentation**	**ours**
CN *vs.* MCI *vs.* AD	0.5802	0.5883	**0.6715**
CN *vs.* MCI	0.6479	0.6856	**0.7436**
sMCI *vs.* pMCI	0.6946	0.7162	**0.7703**
MCI *vs.* AD	0.7965	0.8333	**0.8716**
CN *vs.* AD	0.8898	0.9231	**0.9489**

TABLE 5: Comparison of performance on the multi-class (*i.e.*, CN *vs.* MCI *vs.* AD) classification scenario on the ADNI dataset.

Guidance	**Models**	**mAUC**	**ACC**
	ResNet18 [48]	0.7501 ± 0.046	0.5802 ± 0.041
	SonoNet16 [49]	0.7452 ± 0.069	0.5912 ± 0.056
	VoxCNN [10]	0.7732 ± 0.034	0.5863 ± 0.045
	Liu *et al.* [37]	0.7016 ± 0.056	0.5468 ± 0.069
	Jin *et al.* [38]	0.7294 ± 0.055	0.5901 ± 0.041
	Li *et al.* [42]	0.7559 ± 0.038	0.6115 ± 0.062
✓	Lian *et al.* [40]	0.7671 ± 0.075	0.6257 ± 0.059
	Ours (ResNet18 + XGA)	**0.8123 ± 0.052**	**0.6715 ± 0.051**

for smoothness of the generated map as it enforces each pixel to correlate to its neighbouring pixels. The term of $\mathcal{L}_{\text{map}}$ is responsible for sharpness (or sparsity) of the image as it uses an elastic regularizer. The GAN loss is vital to ensure the overall quality of the counterfactual maps, as seen by the considerable drop in NCC scores when ablating out the GAN loss.

4.2 Diagnostic Model Reinforcement

In this section, we demonstrate the effectiveness of our LEAR framework in reinforcing diagnostic models. To do so, we have divided this section into three parts. First, we consider the CF map transformation, *i.e.*, Eq. (2), as a baseline method for our work. Second, we compare our LEAR framework with state-of-the-art attention methods. Third, we demonstrate the effectiveness of the optimized CF map transformation, *i.e.*, Eq. (20), in improving the quality of visual explanation as well as the performance of diagnostic models.

To verify the effectiveness of our proposed XGA module and its produced guidance map, we compare the diagnostic performance in accuracy (ACC) and multi-class area under the receiver operating characteristic curve (mAUC). Note that we use a five-fold cross-validation setting for all experiments, and use the same indices for all the comparison methods with no data leakage.

4.2.1 Reinforcement via Augmentation

As a baseline method for diagnostic model reinforcement, we utilize the CF map without XGA injection (*i.e.*, Eq. (2)) to produce synthesized images to augment training samples and use those to update the backbone diagnostic model. Specifically, using a CF map defined by Eq. (2), we transformed all train data samples with target labels other than their ground-truth label. For three-class experiments, we produced transformed images with two other target labels. For example, if the input image is an AD subject, we produced NC-transformed and MCI-transformed images for that input image. Finally, those transformed images were used to fine-tune the backbone diagnostic models. We decided to use this augmentation method as a baseline method for our work because it is one of the simplest ways to utilize the CF map in reinforcing the diagnostic performance. To this end, we report the comparison between the backbone, baseline, and our method in Table 4. The improvement (+3.7%) in the classification accuracy of our method over that of the baseline method suggests that our CF maps can indeed capture class-discriminative information and also indicates that these kinds of visual explanation can guide and reinforce a diagnostic model, which supports the motivation behind this study. In the following paragraphs, we will demonstrate that the LEAR framework can further reinforce the diagnostic models with guidance from visual explanation using CF maps.

4.2.2 Comparison to Other Diagnostic Models

To demonstrate the ability of our LEAR framework in reinforcing diagnostic models, we have pre-trained and fixed the weights of a three-class backbone ResNet18 diagnostic model and re-implemented state-of-the-art diagnostic models. To this end, we compare the performances of state-of-the-art *attention-guided* [40], [42] models and *conventional* [10], [37], [38], [48], [49] CNN models in Table 5. Refer to Supplementary S6 for the diagnostic performance of these conventional CNN models applied to our proposed LEAR framework. Notably, they consistently derived performance improvements, thus proving its generalizability.

Notably, our work demonstrates significant improvement over the ResNet18 backbone model (ACC +15.74%) as well as the state-of-the-art CNN models (mean ACC +13.64%). In comparison to conventional CNNs, attention-guided diagnostic models (*e.g.*, Li *et al.* [42] and Lian *et al.* [40]) mostly excel in diagnostic performances. However, these models are guided by *conventional* visual attribution methods, such as CAM, that can only provide coarse-grained guidance. Our LEAR framework, synthesizing and exploiting fine-grained guidance, outperformed all the competing methods by large margins in mAUC and ACC. It is noteworthy that the performance improvements were obtained for all the diagnostic models considered in our experiments (*i.e.*, ResNet18, VoxCNN, and SonoNet16) in the experiments equally. Furthermore, as most of the comparing works were proposed as binary diagnostic models, we have performed a comprehensive binary diagnosis comparison and presented in Supplementary S5.2. Our LEAR framework outperforms all comparing models in all binary class settings (mean ACC: CN *vs.* MCI +14.80%, sMCI *vs.* pMCI +10.82%, MCI *vs.* AD +12.83%, CN *vs.* AD +7.69%).

4.3 Iterative Explanation-Reinforcement Learning

Here, we demonstrate how the iterative learning scheme of our LEAR framework can further improve the diagnostic performances and, thereby, the quality of visual explanation.

4.3.1 Effects in Generalization of a Diagnostic Model

We applied three iterations of our LEAR framework and presented the results in Table 6. In comparison to the back-

TABLE 6: Comparison of performance (ACC) among various iterations on the ADNI dataset.

Setting	ResNet18			
	Backbone	1st	2nd	3rd
CN *vs.* MCI *vs.* AD	0.5802	0.6347	0.6715	0.6715
CN *vs.* MCI	0.6479	0.7014	0.7436	0.7436
sMCI *vs.* pMCI	0.6946	0.7381	0.7703	0.7703
MCI *vs.* AD	0.7965	0.8396	0.8716	0.8716
CN *vs.* AD	0.8898	0.9229	0.9515	0.9489

bone model, the iterations of our LEAR framework have increased the accuracy by +5.49%, +10.64%, and +10.50%, respectively, for each iteration.

For a visual inspection of the changes in class-relevant feature representations, in Fig. 8, we present the CAM visualization along with the CF map over three iterations of EU and RU learning for an AD sample (Subject ID 005_S_0223), which our backbone ResNet18 diagnostic model misclassified with a low class probability. As shown in Fig. 8, the XGA module excels in cases where the confidence of the predictive probability (*i.e.*, values at the bottom of each image) of the backbone diagnostic model is low. Specifically, the CAM obtained from the backbone network vaguely highlights the ventricular region (*i.e.*, the center of an image), whereas the CAM results after iterative learning focus on more meaningful and fine-grained regions of the cortex and hippocampus. Likewise, the first iteration CF map neglected to highlight the hippocampus region (orange box) associated with the AD progression, but the second and third iteration CF maps clearly observed the hypertrophy in the hippocampus region with an equivalent intensity. Thus, the XGA module of our LEAR framework improves the diagnostic model not only in terms of performance but also in terms of innate interpretability of a diagnostic model. Interestingly, we found that diagnostic scores converge after second iteration, but the CAM results continue to qualitatively improve even after the second iteration. The more the iterations that were performed, the more fine-grained and class-discriminative the CAM results became.

4.3.2 Effects in Visual Explanation

First, we selected a CN and AD image from a subject (Subject ID 131_S_0123) whose CF map was of unsatisfying quality. Then, we produced the ground-truth map for CN←AD transformation and CF maps for three iterations of our LEAR framework (Fig. 9). Note that the first iteration CF map does not benefit from reinforcement because the CF map is defined by Eq. (2) at this iteration and is redefined by Eq. (20) from the second iteration onwards.

In the first-iteration CF map, the attribution completely ignores the hypertrophy in the hippocampus (orange box). The attributions in the cortical (green box) and ventricular (purple box) regions are also weak and noisy, making the visual explanation pathologically unreliable. However, the second-iteration CF map successfully captures the hypertrophy in the hippocampus and the attribution to the cortex regions is clearer, but the attribution in the ventricles has become nosier. Finally, the third-iteration CF map clears up the noisy attribution in the ventricles. More diverse results are presented in Supplementary S4.

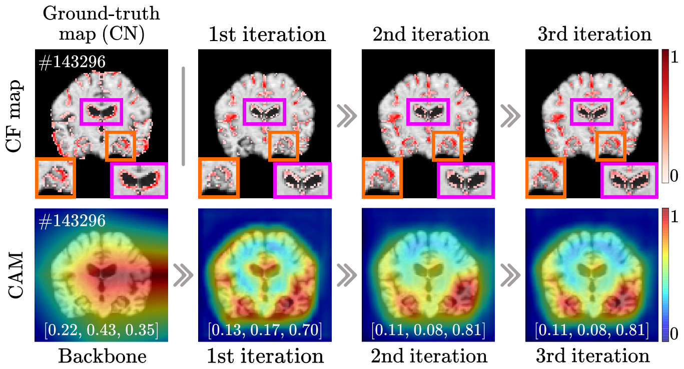

Fig. 8: Counterfactual map and CAM visualization of XGA-injected ResNet18 on the CN *vs.* MCI *vs.* AD scenario with self-iterative training. The values at the bottom of brain images (Subject ID 005_S_0223) are the model's softmax activated logits.

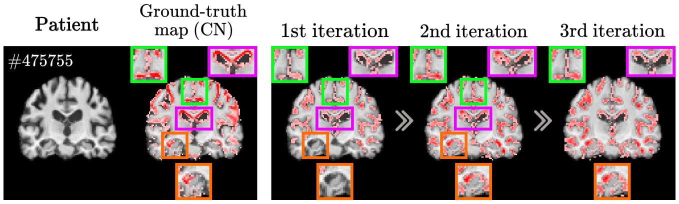

Fig. 9: Reinforced counterfactual map visualization by using iterative optimization on trained ResNet18 (Subject ID 131_S_0123, Image ID on top left corner). The purple, orange, and green boxes correspond to the ventricular, hippocampus, and cortex regions, respectively.

5 Conclusion

With the unprecedented successes of deep learning in various fields, there have been efforts of developing deep-learning methods in medical image analysis including brain disease diagnosis. However, it is still limited for real-world applications due to its unfavorable black-box property.

In this work, we proposed a novel learn-explain-reinforce (LEAR) framework for producing high-quality visual explanations about decision-making in 3D MRI-based AD diagnosis through counterfactual maps generation and for reinforcing a diagnostic model. Specifically, we devised the counterfactual map generator (CMG) to generate *multi-way* counterfactual maps given a pre-trained diagnostic model, an explanation-guided attention (XGA) module for feature representations enhancement, and an iterative reinforcement learning scheme to improve diagnostic performance. Our exhaustive experiments over the ADNI dataset have empirically proved the validity and generalizability of the proposed LEAR framework.

We believe that counterfactual reasoning helps explain a model's decision in an intuitive manner. However, when generating a counterfactual map, it is imperative to reflect other factors such as age, gender, and genes from a causal inference perspective. In that regard, it would be principal research directions to infer causal relations and to learn representations accordingly. The causality-involved learning will make more robust decision-making and better explanations about the decision.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) No. 2022-0-00959 ((Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making) and No. 2019-0-00079 (Department of Artificial Intelligence (Korea University)).

REFERENCES

[1] A. Association *et al.*, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] S. Li, O. Okonkwo, M. Albert, and M.-C. Wang, "Variation in variables that predict progression from MCI to AD dementia over duration of follow-up," *American Journal of Alzheimer's Disease (Columbia, Mo.)*, vol. 2, no. 1, p. 12, 2013.

[3] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[4] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[5] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, S. C. Johnson, A. D. N. Initiative *et al.*, "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," *NeuroImage*, vol. 48, no. 1, pp. 138–149, 2009.

[6] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative *et al.*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[7] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "Supervised discriminative group sparse representation for mild cognitive impairment diagnosis," *Neuroinformatics*, vol. 13, pp. 277–295, 2014.

[8] Y. Shi, H.-I. Suk, Y. Gao, and D. Shen, "Joint coupled-feature representation and coupled boosting for ad diagnosis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2721–2728.

[9] H.-I. Suk and D. Shen, "Deep learning-based feature representation for AD/MCI classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 583–590.

[10] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3D brain MRI classification," in *IEEE International Symposium on Biomedical Imaging*, 2017, pp. 835–838.

[11] H.-I. Suk, S.-W. Lee, D. Shen, Alzheimer's Disease Neuroimaging Initiative *et al.*, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Medical Image Analysis*, vol. 37, pp. 101–113, 2017.

[12] W. Jung, E. Jun, and H.-I. Suk, "Deep recurrent model for individualized prediction of Alzheimer's disease progression," *NeuroImage*, vol. 237, p. 118143, 2021.

[13] H.-I. Suk and D. Shen, "Deep ensemble sparse regression network for Alzheimer's disease diagnosis," in *Machine Learning in Medical Imaging*. Cham: Springer International Publishing, 2016, pp. 113–121.

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[15] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[16] E. Lee, J.-S. Choi, M. Kim, and H.-I. Suk, "Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning," *NeuroImage*, vol. 202, p. 116113, 2019.

[17] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[18] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4942–4950.

[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[20] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *IEEE International Conference on Data Science and Advanced Analytics*, 2018, pp. 80–89.

[21] 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 *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017, pp. 3319–3328.

[23] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[24] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[25] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, 2017, pp. 3145–3153.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[27] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*, 2019, pp. 2376–2384.

[28] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (CaCE)," *arXiv preprint arXiv:1907.07165*, 2019.

[29] P. Wang and N. Vasconcelos, "SCOUT: Self-aware discriminant counterfactual explanations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8981–8990.

[30] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Cybersecurity*, vol. 31, p. 841, 2017.

[31] C. F. Baumgartner, L. M. Koch, K. Can Tezcan, J. Xi Ang, and E. Konukoglu, "Visual feature attribution using wasserstein GANs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8309–8319.

[32] C. Bass, M. da Silva, C. Sudre, P.-D. Tudosiu, S. Smith, and E. Robinson, "ICAM: Interpretable classification via disentangled representations and feature attribution mapping," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[33] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *International Conference on Learning Representations*, 2019.

[34] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021, pp. 650–665.

[35] A. Sauer and A. Geiger, "Counterfactual generative networks," in *International Conference on Learning Representations*, 2021.

[36] S. Dash, V. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 915–924.

[37] S. Liu, C. Yadav, C. Fernandez-Granda, and N. Razavian, "On the design of convolutional neural networks for automatic detection of Alzheimer's disease," in *Machine Learning for Health Workshop*. PMLR, 2020, pp. 184–201.

[38] D. Jin, J. Xu, K. Zhao, F. Hu, Z. Yang, B. Liu, T. Jiang, and Y. Liu, "Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration," in *IEEE International Symposium on Biomedical Imaging*, 2019, pp. 1047–1051.

[39] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, "An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Journal of Biomedical and Health Informatics*, vol. 14, no. 8, 2021.

[40] C. Lian, M. Liu, Y. Pan, and D. Shen, "Attention-guided hybrid network for dementia diagnosis with structural MR images," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 1992–2003, 2020.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[42] Q. Li, X. Xing, Y. Sun, B. Xiao, H. Wei, Q. Huo, M. Zhang, X. S. Zhou, Y. Zhan, Z. Xue *et al.*, "Novel iterative attention focusing strategy for joint pathology localization and prediction of MCI progression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 307–315.

[43] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K. N. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, and K. Bae, "Explaining convolutional neural networks through attributionbased input sampling and block-wise feature aggregation," in *AAAI Conference on Artificial Intelligence*, 2021.

[44] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[45] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Advances in Neural Information Processing Systems*, 2018, pp. 592–603.

[46] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, 2017, pp. 6967–6976.

[47] Y. Wang, H. Su, B. Zhang, and X. Hu, "Learning reliable visual saliency for model explanations," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1796–1807, 2019.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[49] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, "SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[50] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 880–893, 2018.

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[54] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[55] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[56] S. Gottlieb and C.-W. Shu, "Total variation diminishing Runge-Kutta schemes," *Mathematics of Computation*, vol. 67, no. 221, pp. 73–85, 1998.

[57] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Learning what and where to attend," in *International Conference on Learning Representations*, 2019.

[58] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[59] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[60] Y. LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[61] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*, 2018, pp. 2649–2658.

[62] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869 – 877, 2005.

[63] B. C. Dickerson, I. Goncharova, M. Sullivan, C. Forchetti, R. Wilson, D. Bennett, L. A. Beckett, and L. deToledo Morrell, "MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease," *Neurobiology of Aging*, vol. 22, no. 5, pp. 747–754, 2001.

[64] Y. Fan, N. Batmanghelich, C. M. Clark, C. Davatzikos, A. D. N. Initiative *et al.*, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," *NeuroImage*, vol. 39, no. 4, pp. 1731–1743, 2008.

[65] E. Gerardin, G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehéricy, L. Garnero *et al.*, "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging," *NeuroImage*, vol. 47, no. 4, pp. 1476–1486, 2009.

[66] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, and P. Kickingereder, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human Brain Mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[67] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782 – 790, 2012.

Kwanseok Oh received the BS degree in Electronic Control and Engineering from Hanbat National University, Daejeon, South Korea, in 2020. He is currently pursuing the PhD degree with the Department of Artificial Intelligence, Korea University, Seoul, South Korea.

His current research interests include explainable AI, computer vision, and machine/deep learning.

Jee Seok Yoon received the BS degree in Computer Science and Engineering from Korea University, Seoul, South Korea, in 2018. He is currently pursuing the PhD degree with the Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea.

His current research interests include explainable AI, computer vision, and representation learning.

Heung-Il Suk received the BS and MS degrees in computer engineering from Pukyong National University, Busan, Korea, in 2004 and 2007, respectively, and the the PhD degree in computer science and engineering from Korea University, Seoul, South Korea, in 2012.

From 2012 to 2014, he was a Post-Doctoral Research Associate with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently an Associate Professor with the Department of Artificial Intelligence and the Department of Brain and Cognitive Engineering, Korea University. He was awarded a Kakao Faculty Fellowship from Kakao and a Young Researcher Award from Korean Society for Human Brain Mapping (KHBM) in 2018 and 2019. His research interests include machine/deep learning, explainable AI, biomedical data analysis, and brain-computer interface.

Dr. Suk serves as an Editorial Board Member for Electronics, Frontiers in Neuroscience, International Journal of Imaging Systems and Technology (IJIST), and a Program Committee or a Reviewer for NeurIPS, ICML, ICLR, AAAI, IJCAI, MICCAI, AISTATS, *etc.*