IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 27, 2021, accepted October 10, 2021, date of publication October 19, 2021, date of current version October 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3121609

Longitudinal Brain MR Image Modeling Using Personalized Memory for Alzheimer's Disease

SEONG TAE KIM[1], (Member, IEEE), UMUT KÜÇÜKASLAN[2], AND NASSIR NAVAB[2,3], (Senior Member, IEEE)
[1]Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea
[2]Chair for Computer Aided Medical Procedures, Technical University of Munich, 80333 München, Germany
[3]Chair for Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, MD 21218, USA

Corresponding author: Seong Tae Kim (st.kim@khu.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (Development of Provisional Intelligence Based on Long-term Visual Memory Network) under Grant 2020-0-00004, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) under Grant 2021R1G1A1094990.

ABSTRACT Longitudinal analysis of a disease is an important issue to understand its progression and design prognosis and early diagnostic tools. From the longitudinal images where data is collected from multiple time points, both the spatial structural information and the longitudinal variations are captured. The temporal dynamics are more informative than static observations of the symptoms, particularly for neurodegenerative diseases such as Alzheimer's disease, whose progression spans over the years with early subtle changes. In this paper, we propose a new generative framework to predict the lesion progression over time. Our method first encodes images into the structural and longitudinal state vectors, where interpolation or extrapolation of feature vectors in the time axis can be performed for the manipulation of these feature vectors. These processed feature vectors can be decoded into image space to predict the image at the time point which we are interested in. During the training, we force the model to encode longitudinal changes into longitudinal state features and capture the structural information in a separate vector. Moreover, we introduce a personalized memory for the online update scheme, which adapts the model to the target subject, which helps the model preserve fine details of brain image structures in each subject. Experimental results on the public longitudinal brain magnetic resonance imaging dataset show the effectiveness of the proposed method.

INDEX TERMS Brain MR images, deep learning, generative model, longitudinal analysis, personalized prediction, memory network.

I. INTRODUCTION

Longitudinal analysis of a disease takes scans of patients at different time points where structural abnormalities and temporal changes are captured. The changes of anomalies over time can be more informative than the static information for certain disease types such as neurodegenerative diseases, whose progression span over the years with early subtle changes [19]. For example, the precursors of Alzheimer's disease are present earlier than the first symptoms of the disease. In addition, the change in certain biomarkers seems to predict Alzheimer's disease before the first clinical symptoms are present [3]. Magnetic Resonance (MR) images determine atrophy, and hippocampal volume is a good predictor for conversion from mild cognitive impairment to Alzheimer's disease [6]. Also, lateral ventricular growth is a distinct feature in Alzheimer's disease and can assess disease progression [20]. This structure tends to grow in the axial slices of the brain MR image in patients developing Alzheimer's [20]. Therefore, the disease progression can easily be assessed using longitudinal structural MR images. An interesting problem here is to estimate the future conditions of patients using their previous MRI scans. Predicting future slices of patients can help medical doctors to assess the disease progression speed and provide proper treatment to patients [4].

Autoencoders [10], [17], [29] are known that they can learn a manifold where the input image is mapped to the

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

dense and lower-dimensional latent features by the encoder. Then, the decoder learns to reconstruct the image from the latent features. From this manifold, new samples can be generated via interpolations or more complex sampling schemes between latent features and decoding of it to the target image space if the latent space is smooth. Lee *et al.* use an encoder-decoder structure to generate realistic lesions by using feature manipulation in the latent space. They show interpolation and extrapolation in the latent space to generate new samples [17]. Louis *et al.* use a recurrent neural network to predict the parameters of a disease progression model, which is used to sample vectors corresponding to different time points [19]. Pathan *et al.* propose a decoder to generate a vector field that is used to deform input images to generate output images, in which the underlying changes in successive images are captured [21]. Bowles *et al.* propose a Wasserstein generative adversarial network to model brain MR images with Alzheimer's disease features. This allows for synthetic images based upon an individual subject's MR image to be produced, expressing different levels of the features associated with the disease [4].

In this paper, we propose a new model that generates patient-specific longitudinal brain images by using personalized memory. Based on two reference scans taken at different time points of the target patient, our model can generate a scan at any target time point by modeling the patient-specific progression in the personalized memory. To better model the latent features of brain images, we separate the latent feature vectors to the structural features and longitudinal state features. During training, the structure vectors of longitudinal sequences from the same patients are encouraged to be close. The temporal changes over time are encoded in the temporal state features. In addition, to improve the quality of the generated images, we devise a personalized memory with the online adaptive training where the model is fine-tuned to the target patient in a short time. The contributions of this paper can be summarized as

- The proposed method can effectively model the temporal changes of longitudinal brain MR images by separating the structural features and the temporal state features.
- The online adaptive training scheme to construct the personalized memory is devised, which can significantly improve the quality of generated images at the test time.
- Comparative experiments have been conducted to show the effectiveness of the proposed method. Experimental results show that the proposed adaptive training scheme for personalized modeling of longitudinal progression in the memory can be a promising solution for the future prediction of brain MR images.

II. RELATED WORK

Longitudinal medical image sequences consist of multiple images taken at different time points from the same patient [16], [18], [24]. Generative models can be relevant for predicting missing and future slices in a longitudinal image sequence.

Recently, the generative models which can learn the underlying data distribution have been widely explored. Radford *et al.* propose a deep convolutional generative adversarial network (GAN) which consists of generator and discriminator [22]. Both generator and discriminator are trained together in an adversarial way, which makes it possible to generate perceptually meaningful and smooth images. Kingma *et al.* [15] propose a new generative model named GLOW using normalizing flows [8], [25]. Furthermore, these studies [15], [22], [26] show that perceptually meaningful manifold can be mapped into a linear path in the latent space, and we can generate new images by feature manipulation. In other words, it is possible to generate an output image in the desired way.

In the medical domain, Bowles *et al.* [4] use a Wasserstein GAN (WGAN) to reconstruct brain MR images from the random vector. They assumed that the discriminator is a Lipschitz function, and it is encouraged to be in a compact space by clipping the gradient values during the training. However, the WGAN can only map latent vectors into images, and they used a gradient descent method on input to find the latent vector from initialized one. [4] computed the average latent vectors for three different groups of Alzheimer's disease (AD) patients, mild cognitive impairment (MCI) patients, and conditionally normal (CN) subjects. The limitation of the method is that the progression modeling is not patient-specific.

Pathan *et al.* [21] use a vector field which is used to deform input scan to generate the target scan at different time points. The diffeomorphic maps are estimated by the large deformation diffeomorphic metric mapping (LDDMM) framework for the vector field. In other words, predicting MR images at different time points is modeled as the problem of finding the corresponding deformation field. The input for LDDMM is latent features encoded by long-short term memory (LSTM) networks. Louis *et al.* [19] devise a framework where the model can learn a low-dimensional space of disease progression with respect to time. The CNN is used to extract image features, and the recurrent neural network (RNN) is used to predict disease progression parameters. However, the result images lack the fine details of the brain. Lee *et al.* [17] propose an autoencoder network to generate realistic lesions. The linear interpolation or extrapolation of two latent vectors is used to generate new lesions in the latent space. As a result, the smooth manifold is learned in the latent space, which makes it possible to generate realistic lesions.

Contrary to the existing generative models, in this paper, we propose the disentanglement of structure features and longitudinal state features. In addition, we firstly propose a personalized memory through an online update to better model the longitudinal progression of the patient at the test time.

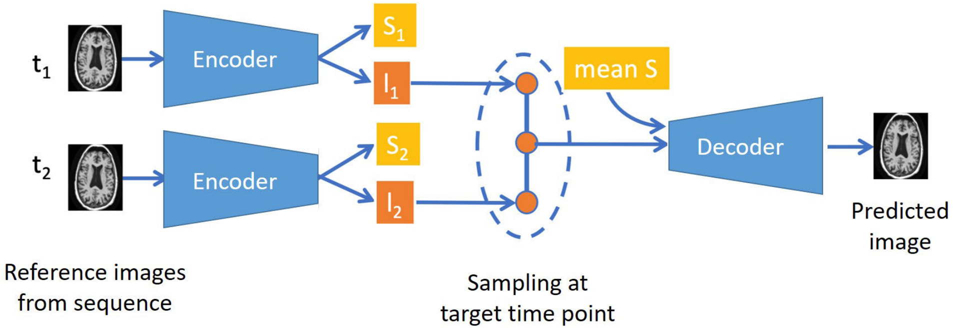

FIGURE 1. The overall framework of the proposed method. The proposed model consists of an encoder and a decoder where the latent space learned by the encoder is used to manipulate the latent feature vector. The latent feature vector is sampled and decoded to predict the image at the target time point.

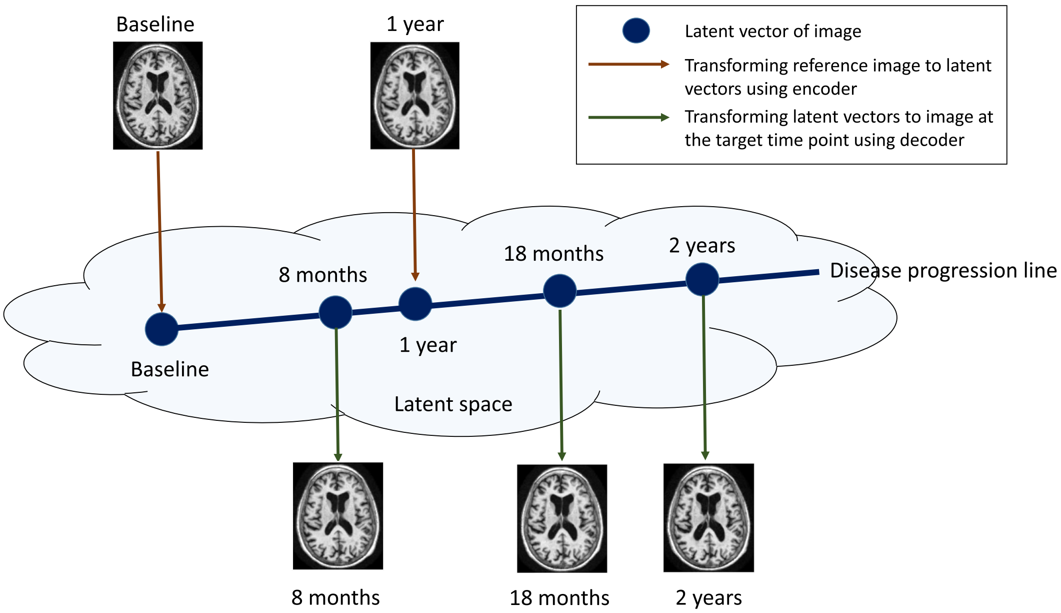
Baseline
1 year
Latent vector of image
Transforming reference image to latent vectors using encoder
Transforming latent vectors to image at the target time point using decoder
8 months
18 months
2 years
Disease progression line
Baseline
1 year
Latent space
8 months
18 months
2 years

FIGURE 2. The disease progression in the latent space is encouraged to be linear, therefore the encodings of images from the longitudinal sequence of a patient align on a line. By using the latent vectors of two reference images, the other latent vectors can be derived, which can be further used for predicting images.

III. METHOD

In this paper, we develop a model that generates patient-specific (i.e., personalized) missing and future MR images using two reference images. For this purpose, we propose an autoencoder network that first encodes images to latent vectors, then decodes latent vectors back to image space. In other words, two reference images, which are from different time points and show different stages of the disease, are firstly encoded to latent vectors. Then, the target time point image's latent vectors are calculated by interpolation or extrapolation using the reference image's latent vectors. For each longitudinal sequence and each two image pair that can be chosen within that sequence, we choose those two images as the reference and generate the brain MR image at the target time point.

To better model the structure of the brain and temporal changes over time, we devise the proposed model as follows. Our model encodes input images into two vectors in the bottleneck layer: *structure feature* and *longitudinal state feature*. The structure feature is responsible for encoding the brain structures in the image that are stationary. In contrast, the longitudinal state feature is responsible for encoding the states of the structures that are changing over time. The decoder reconstructs the output image using the structure data encoded in the structure feature and the states of the changing structures encoded in the longitudinal state feature as in Figure 1. The structure features of the reference images are averaged for decoding brain images at the target time point.

A. FEATURE MANIPULATION IN LATENT SPACE FOR IMAGE PREDICTION

We train our model to learn a latent space where the disease progression is mapped to a linear path in the latent space, and the position of latent vectors on the line is a function of time as can be seen in Figure 2. We can model the disease progression by training this latent space by using the two encoded latent vectors. And the latent vector of the image corresponding to any target time point can be computed via linear interpolation or extrapolation of the encoded latent vectors of two reference images.

Let $\{t_i, \boldsymbol{u}_i\}_{i=1}^{N}$ be a longitudinal image sequence where $\boldsymbol{u}_i$ denotes the MR image at the time t_i. N denotes the number of images in the sequence. Then the encoder network maps input image $\boldsymbol{u}_i$ to the structure feature $\boldsymbol{s}_i$ and the longitudinal state feature $\boldsymbol{l}_i$ as $E(\boldsymbol{u}_i) = [\boldsymbol{s}_i, \boldsymbol{l}_i]$. Also, let $D(\cdot)$ be a function, that represents the decoder network, where it maps structure feature $\boldsymbol{s}_i$ and longitudinal state feature $\boldsymbol{l}_i$ to the image $\boldsymbol{u}_i$ as $D(\boldsymbol{s}_i, \boldsymbol{l}_i) = \boldsymbol{u}_i$.

To encode the disease progression on a line in the latent space where we can manipulate the latent features with the longitudinal state feature, we enforce the model that the structural features of the images from the same patients be the same. This assumption constrains the network to encode the temporal changes into the longitudinal state features and brain structures common within images in the longitudinal sequence into the structure features. In other words, a longitudinal image sequence can be represented with the patient-specific structure features $\boldsymbol{s}_c$ and associating longitudinal state features $\{\boldsymbol{l}_i\}_{i=1}^{N}$ by the encoder:

$$\{t_i, \boldsymbol{u}_i\}_{i=1}^{N} \underset{\text{decoder}}{\overset{\text{encoder}}{\rightleftarrows}} \{\boldsymbol{s}_i, \boldsymbol{l}_i\}_{i=1}^{N}, \tag{1}$$

where $\boldsymbol{s}_i = \boldsymbol{s}_c$ for all i.

The sampling in the latent space is performed using a linear model of disease progression as in Figure 2, which means the difference vectors among latent vectors are proportional to relative time differences in the longitudinal sequence. To learn the linear space with respect to time, we introduce a new training scheme which will be explained in Subsection III. B. As a result, we can predict brain images at any time point on the disease progression line using two latent vectors.

Since the structural features are the same for all samples within a longitudinal sequence, it is enough to manipulate longitudinal state features. Let $(t_1, \boldsymbol{u}_1)$ and $(t_2, \boldsymbol{u}_2)$ be two reference samples that are from the same patient at the different time points, and $(\boldsymbol{s}_1, \boldsymbol{l}_1)$ and $(\boldsymbol{s}_2, \boldsymbol{l}_2)$ be the latent features of the corresponding samples. Then we can generate the latent features $\boldsymbol{s}_x$ and $\boldsymbol{l}_x$ at the target time point t_x as

$$\boldsymbol{s}_x = \boldsymbol{s}_1 = \boldsymbol{s}_2, \tag{2}$$

$$\boldsymbol{l}_x = \frac{t_x - t_1}{t_2 - t_1}(\boldsymbol{l}_2 - \boldsymbol{l}_1). \tag{3}$$

Finally, the image at the target time point t_x is predicted as

$$D([\boldsymbol{s}_x, \boldsymbol{l}_x]) = \hat{\boldsymbol{u}}_{\boldsymbol{x}}. \tag{4}$$

In practice, we use the mean structure features calculated from $(\boldsymbol{s}_1, \boldsymbol{s}_2)$ for sampling.

B. NETWORK TRAINING

To train the proposed network, we construct training data consisting of longitudinal sequences of three images from the same patient at different time points and the image timestamps. Each image is predicted using the other two images by setting the other two images as the reference images in each sequence. The total loss is the summation of the three losses calculated from the prediction of three images in the sequence in an iterative way. The loss for each sample has two contributing parts: *target image reconstruction loss* and *structure feature loss*. The target image reconstruction loss computes the mean squared error between the predicted image and the ground truth image to ensure that the reconstruction is close to the target scan (real image at the target time point). In contrast, the structure feature loss computes the mean squared error between the encoded structure features between the two reference images. The structure feature loss constrains the deep network to encode the disease progression only in the longitudinal state features.

In detail, let $\{t_x, \boldsymbol{u}_x\}$ be the target sample that the image at time point t_x. Then $\boldsymbol{u}_x$ is generated using the other reference images $\{t_1, \boldsymbol{u}_1\}$ and $\{t_2, \boldsymbol{u}_2\}$ in the sequence as described in subsection III.A. Therefore, the loss for predicting the sample $\{t_x, \boldsymbol{u}_x\}$ from the reference samples $\{t_i, \boldsymbol{u}_i\}_{i \neq x}$ is as follow:

$$L(\{t_x, \boldsymbol{u}_x\}) = \underbrace{\sum |\boldsymbol{u}_x - \hat{\boldsymbol{u}}_x|^2}_{\text{Target Image Reconstruction Loss}} + \underbrace{w \sum |\boldsymbol{s}_1 - \boldsymbol{s}_2|^2}_{\text{Structure Feature Loss}}, \tag{5}$$

where w is the weight for balancing two loss functions.

C. PERSONALIZED MEMORY WITH ONLINE ADAPTIVE TRAINING

The strength of our method is that it can generate sharp and detailed images, and the disease progression is well-captured in the generated longitudinal sequence, which is boosted by the online adaptive training in the personalized memory. The network trained on a training dataset is capable of capturing the temporal changes, i.e., the growth of the lateral ventricles, as well as the general structure in the reference images. But it lacks reconstructing the fine details of the complex brain images in each patient. In particular, the brain folds and the shape of the lateral ventricles in the predicted image are slightly different from those in the reference images.

To overcome this limitation, we propose an online adaptive training method as shown in Fig. 3. The proposed method updates the trained model to the personalized memory for the target patient by using the reference images. Training the network for a few steps to make it personalized for the given patient makes the deep network generate sharper and more detailed images and captures the disease progression better.

The training scheme is similar to the training explained in Subsection II.B, but the difference is that a target image is unavailable. Instead, we use the reference images for image

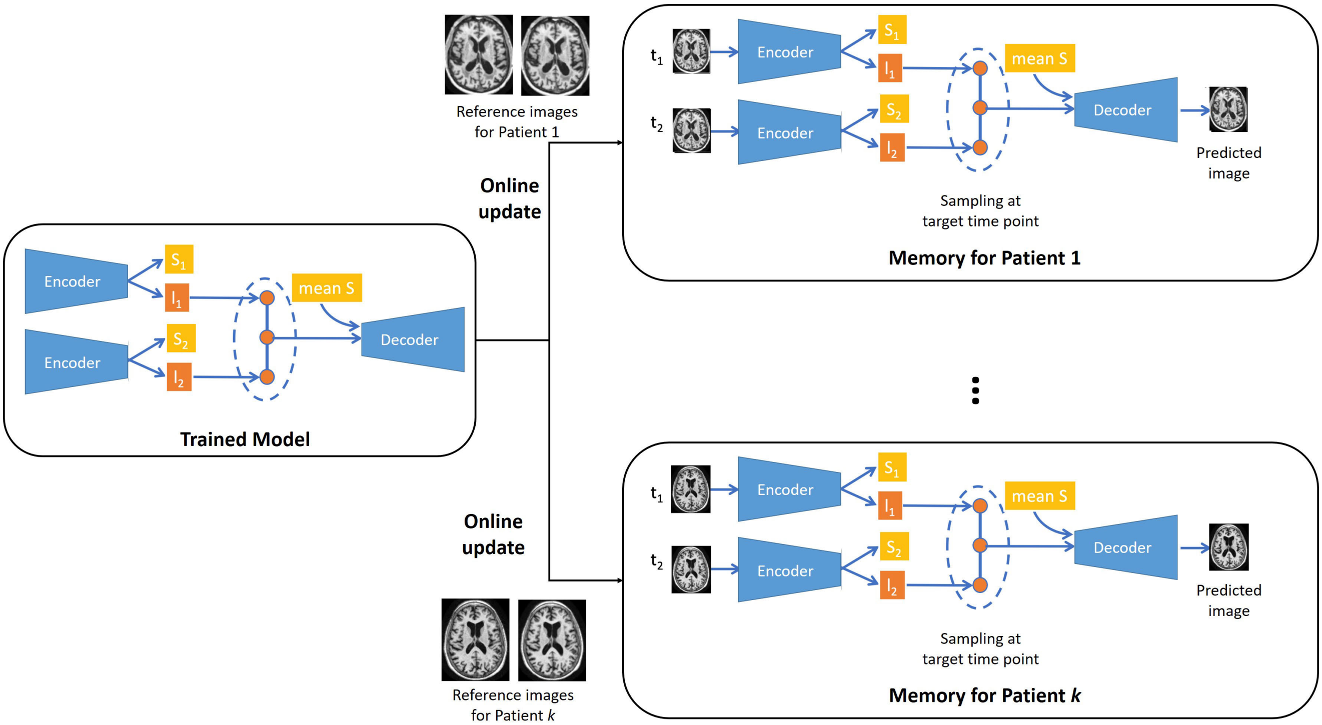

FIGURE 3. The proposed personalized memory with online update at the test time. The trained model is shortly updated to better model the progression in the personalized memory at the test time based on the reference images from the target patient.

reconstruction loss as

$$L_{online} = |\boldsymbol{u}_1 - \hat{\boldsymbol{u}}_1|^2 + |\boldsymbol{u}_2 - \hat{\boldsymbol{u}}_2|^2 + w|\boldsymbol{s}_1 - \boldsymbol{s}_2|^2. \qquad (6)$$

IV. EXPERIMENTS

A. DATASET

To verify our method, we use the public ADNI dataset [1], which includes longitudinal brain MRI scans of three different groups of patients: Alzheimer's disease (AD), Mild cognitive impairment (MCI), and Cognitively normal (CN). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, position emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.[1] We select dataset (ADNI1:Complete 2Yr 1.5T [1]) which includes screening, 6 months, 1 year, 18 months (MCI only), and 2 years scans for each patient. To extract axial slices from each scan, we first register the follow-up scans of a patient to its first scan using 3D rigid-body registration using nibabel[2] and DIPY[3] libraries. The rigid-body registration is performed by applying the center of mass transform, translation transform, and rigid body transform successively to the follow-up scan. The mutual information is used as the metric with a multi-level optimization scheme.

TABLE 1. Statistics of the dataset used in this study (ADNI1:Complete 2Yr 1.5T [1]). The numbers represent the number of patients (and the number of extracted images).

Patient group	Train	Validation	Test	Total
AD	62 (596)	7 (72)	18 (192)	87 (860)
MCI	134 (2,676)	25 (500)	34 (680)	193 (3,856)
CN	106 (1,268)	13 (140)	37 (440)	156 (1,848)
Total	302(4,540)	45(712)	89(1,312)	436(6,564)

We extract four axial slices showing lateral ventricles from each volume and form four longitudinal image sequences for each patient. First, we choose the location of slices in the image, then extract from all scans, reference, and registered follow-up scans. Table 1 shows the dataset statistics for the training, validation, and test set used in this study.

Voxel intensity is normalized after clipping outlier values. To normalize the image range, we calculate the average μ and standard deviation σ of pixels on the brain in each slice, separately. The background is not included in this calculation. Then, the outlier values which are bigger than $\mu + 1.8 \times \sigma$ are clipped, and a min-max normalization is conducted from the clipped image. As a result, we obtain a dataset that consists of registered longitudinal axial slice images, as shown in Figure 4.

B. IMPLEMENTATION DETAILS

For training the model, ADAM optimizer [14] is used with a learning rate of 0.0001 and weight decays of $\beta_1 = 0.5$

[1] For up-to-date information, see www.adni-info.org
[2] https://nipy.org/nibabel/
[3] https://dipy.org/

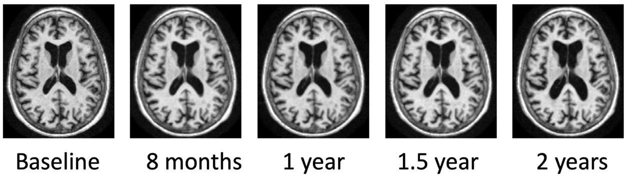

FIGURE 4. Longitudinal axial brain MRI slices of a mild cognitive impairment patient from ADNI dataset [1]. We extracted four such sequences for each patient, each showing a different part of the lateral ventricle.

and $\beta_2 = 0.999$ for all models. For online adaptive training, we use two reference images of patients, which makes the model learn the shapes of the structures in the personalized memory.

The whole brain is resized to 64 × 64 and it is used as an input for the model in this study. For the encoder, four convolutional layers with the filter size of 3 with a stride of 2 are used. Each convolutional layer is followed by Swish activation function [23]. Swish activation is used in this study because it shows stable and superior performances for various tasks including image construction and restoration tasks [9], [27], [28]. The number of filters in the convolutional layers are 64, 128, 256, 512, respectively. The size of the input is halved at each layer. The output of the final convolutional layer is reshaped into a vector, which is transformed into two vectors of *structure vector* and *longitudinal state vector* by using separate fully-connected layers. For the structure vector, the hyperbolic tangent function is used as the activation function. The size of structural features and longitudinal state feature are 100 and 1, respectively.

For the decoder, the inputs are structure feature vector and longitudinal state vector, each of which is transformed to a higher dimensional vector using a fully connected layer. First, two feature vectors are aggregated by adding them, and then the feature vector is reshaped into a 3D feature map. The reshaped 3D feature map is first upsampled and then processed with a 2D convolutional layer with Swish activation [23]. The upsampling operation doubles the height and width of the feature map. The upsampling layer, 2D convolutional layer, and activation layer are repeated four times in total. The filter size of 3 with a stride of 1 is used for the convolutional layers in the decoder. The number of filters in the convolutional layers are 256, 128, 64, 1, respectively. The Swish activation function [23] is used for the first three convolutional layers, and the sigmoid activation function is used for the last convolutional layer. Figure 5 shows the detailed structure of the encoder and the decoder used in this study. The code is publicly available.[4]

The number of training steps in online adaptive training is empirically set to 100, which takes around 20 seconds per patient on an Nvidia Tesla T4 GPU. The effect of training steps in the online adaptive training will be introduced in Subsection V.B.

TABLE 2. Summary of method with different architectures and the training configurations used in this paper. S.L., Dis., P.M. denote sequence learning, disentanglement of structure and longitudinal state features, and personalized memory, respectively.

Method	S.L.	Dis.	P.M.
Proposed method (+ personalized memory)	✓	✓	✓
Proposed method (+ disentanglement)	✓	✓	
Autoencoder (+ sequence learning)	✓		
Autoencoder			

C. EVALUATION

We evaluate the quality of the predicted images by using Structural Similarity Index (SSIM) [30] and mean squared error (MSE) on the independent test data. For each longitudinal sequence, we predict target images by using two reference images and compute the evaluation score between the predicted and ground truth images. It includes predicting previous, missing, and future scans.

D. COMPARISON

For comparison, we implement different methods for encoding latent features as follows:

- **Wasserstein GAN** [4]: Reimplementation of [4].
- **Autoencoder** [17]: This method is implemented based on [17]. The autoencoder is trained for each sample by using the reference image reconstruction loss.
- **Autoencoder (+ sequence learning)**: It is an extension of baseline *autoencoder* where the model is trained with the target image reconstruction loss by considering the longitudinal sequence.
- **Proposed method (+ disentanglement)**: It encodes the structure features and the longitudinal state features separately, as in Figure 1. The model is trained by the target image reconstruction loss and the structure feature loss in Eq. (4).
- **Proposed method (+ personalized memory)**: It uses personalized memory with online adaptive training at the test time. In other words, the prediction is conducted from the model which is trained by Eq. (4) and Eq. (5).

Table 2 shows the summary of the method with different architectures and the training configurations. After training the model, at the test time, Autoencoder [17], Autoencoder (+ sequence learning), Proposed method (+ disentanglement), Proposed method (+ personalized memory) manipulate (e.g., interpolation or extrapolation) the two feature vectors from the reference images and decode the image of the target time as described in III.A. Please note that the Wasserstein GAN model does not provide a suitable way to predict images at the target time points. For that reason, we reconstruct images themselves from [4] (not a prediction but a reconstruction) and calculate the evaluation metric to show the general image quality of Wasserstein GAN in this study.

[4] https://github.com/umutkucukaslan/longitudinalMR

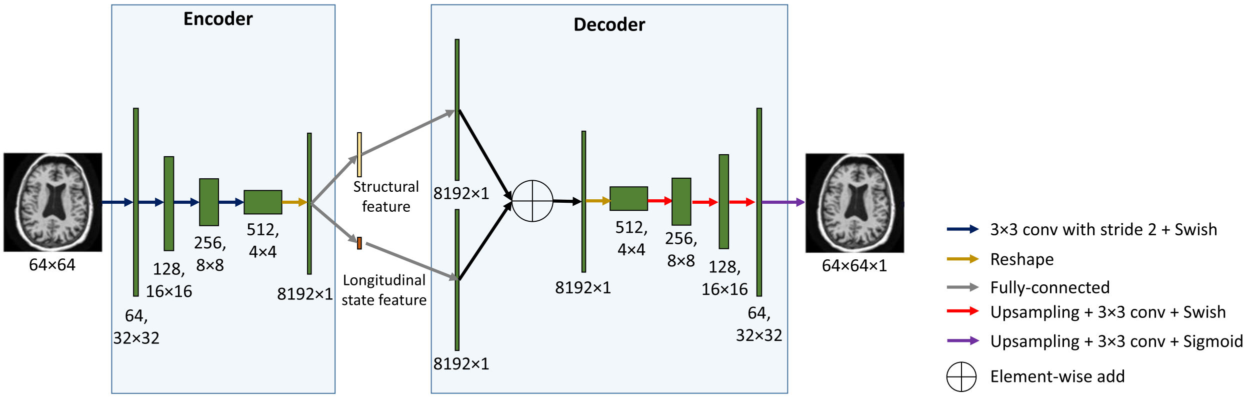

FIGURE 5. **Overall structure of the encoder and the decoder used in this study.**

V. RESULTS

A. ABLATION STUDIES

In this paper, we propose a new model with sequence learning, disentanglement of structure features and longitudinal state features, and personalized memory with online adaptive training. Firstly, we evaluate the effectiveness of each module for predicting longitudinal brain MR images. Table 3 shows the results of the ablation study where we compare autoencoder, autoencoder (+ sequence learning), proposed method (+ disentanglement), and the proposed method (+ personalized memory). SSIM and MSE are calculated on the test set. As shown in the table, autoencoder (+ sequence learning), which uses the targe image reconstruction loss, improves the SSIM (from 0.528 to 0.535) and MSE (from 0.0272 to 0.0266) compared to the baseline autoencoder. The proposed method (+ disentanglement) achieves SSIM of 0.543±0.060 and MSE of 0.0301±0.0071. The improvement is statistically significant compared to autoencoder (+ sequence learning) and baseline autoencoder (p<0.05 by t-test [2]). Furthermore, by using personalized memory based on an online adaptive training scheme, the proposed method achieves SSIM of 0.953±0.015 and MSE of 0.0041±0.0020. The performance improvement is statistically significant compared to the proposed method (+disentanglement) (p<0.01). Based on the ablation studies, we verify the effectiveness of each module, and in particular, the effect of using personalized memory is significant.

TABLE 3. **Ablation study on the test set. The statistics of SSIM and MSE scores for different methods are reported. The average and standard deviation are reported. * denotes the case that the difference with the proposed method is statistically significant (p<0.05).**

Method	SSIM	MSE
Proposed method (+ personalized memory)	**0.953 ± 0.015**	**0.0041 ± 0.0020**
Proposed method (+ disentanglement)	0.543 ± 0.060	0.0301 ± 0.0071
Autoencoder (+ sequence learning)	0.535 ± 0.057*	0.0266 ± 0.0058*
Autoencoder	0.528 ± 0.063*	0.0272 ± 0.0067*

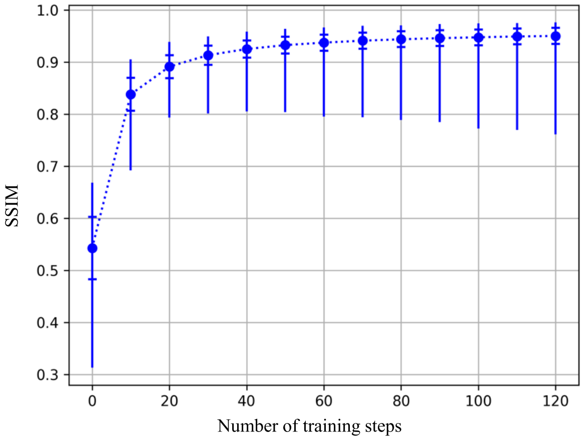

FIGURE 6. **SSIM with respect to the number of training steps in the online adaptation.**

B. EFFECT OF NUMBER OF TRAINING STEPS IN ONLINE ADAPTIVE TRAINING

In this subsection, the number of training steps for online adaptation is investigated. For this purpose, we measure the SSIM with respect to the number of training steps in every 10 training steps in the range of [0, 120]. The number of training steps of 0 denotes the model without online adaptation, which is only trained with the loss function (Eq. 4). Figure 6 shows SSIM with respect to the number of training steps. As shown in the figure, the SSIM is significantly increased when the online adaptation is applied. The biggest improvement is observed when the number of training steps is changed from 0 to 10. The SSIM is increased as the number of training steps increases but it seems to be saturated and it is not sensitive near the number of training step 100. Therefore, we set the number of training steps as 100.

C. COMPARISON WITH OTHER METHODS

In this subsection, we compare the proposed method with other approaches [4], [17]. Table 4 shows the SSIM and

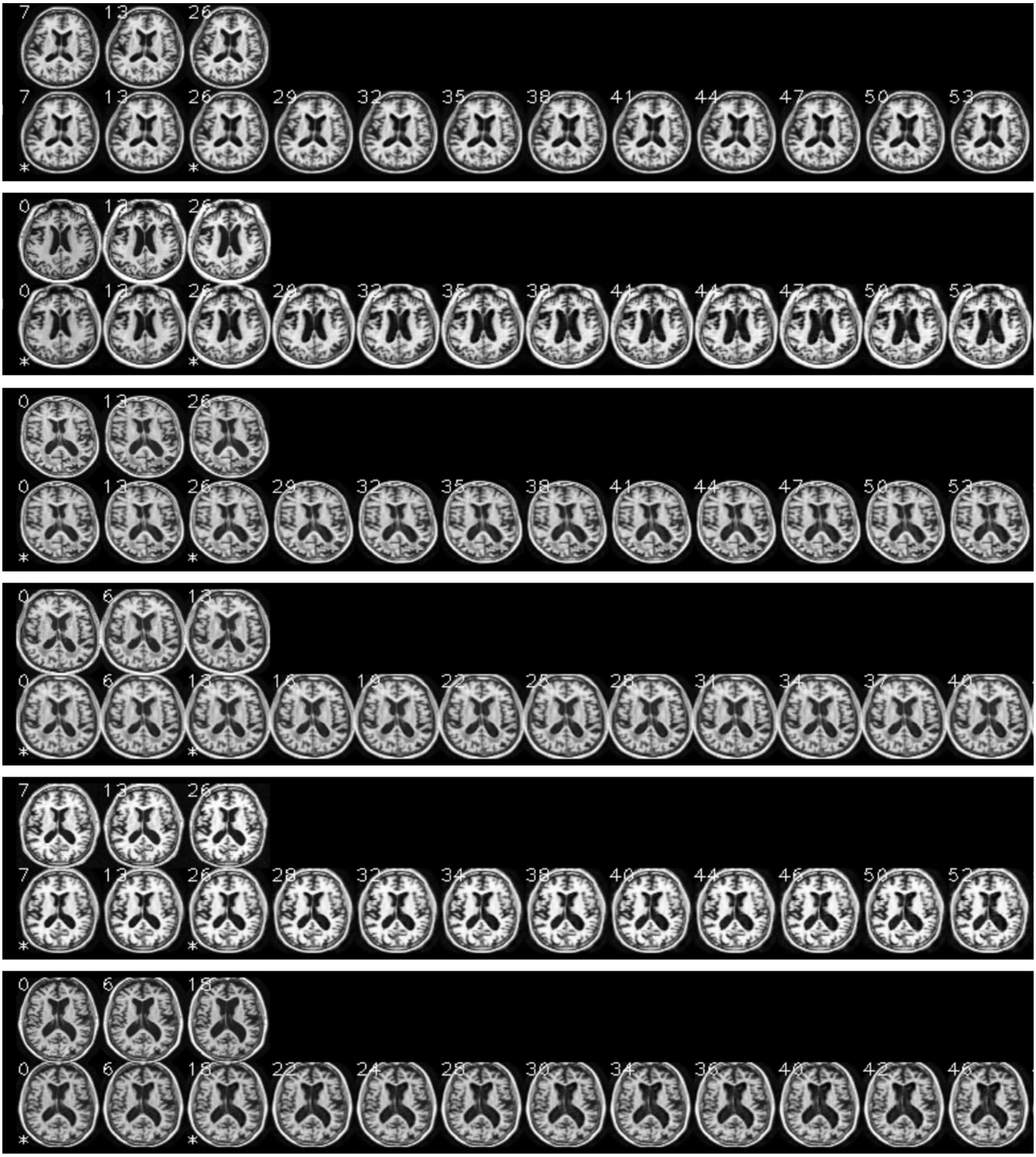

FIGURE 7. Examples of generated sequences using two reference images. Stars show the reference images. Upper left numbers show relative months. The first row shows ground truth images, whereas the second row shows the generated sequence.

MSE measured on the test set from the different methods. Lee *et al.* [17] achieves SSIM of 0.528 ± 0.063 and MSE of 0.0272 ± 0.0067. Please note that Wasserstein GAN model (Bowles *et al.* [4]) does not provide a suitable way to predict images at the target time points. For that reason, we evaluate the reconstruction quality instead of prediction quality. We reconstructed images from their encoded latent features themselves without any manipulation with respect to time information, which enables us to examine image quality degradation just due to the encoding-decoding procedure, and

TABLE 4. Comparison with other methods on the test set. The statistics of SSIM and MSE scores for different methods are reported. The average and standard deviation are reported. * denotes the case that the difference with the proposed method is statistically significant (*p*<0.05).

Method	SSIM	MSE
Proposed method	**0.953 ± 0.015**	**0.0041 ± 0.0020**
Bowles et al. [4]	0.679 ± 0.058*	0.0198 ± 0.0055*
Lee et al. [17]	0.528 ± 0.063*	0.0272 ± 0.0067*

sets an upper bound for SSIM and MSE scores for the target time point image predictions. Bowles *et al.* [4] achieves SSIM of 0.679 ± 0.058 and MSE of 0.0198 ± 0.0055. It cannot predict the patient-wise target timepoint prediction, and the quality will further decrease if it conducts prediction, not reconstruction. Compared to these two approaches, the proposed method can achieve SSIM of 0.953 ± 0.015 and MSE of 0.0041±0.0020, which significantly outperforms previous approaches [4], [17] ($p<0.01$).

D. VISUAL RESULTS

Figure 7 shows example sequences predicted by using the proposed method (+ personalized memory). As shown in the figure, the proposed method can predict the MR images with high quality at the target time point.

E. DISCUSSION

In this study, we propose a new method to model the longitudinal progression of brain MR images by using personalized memory. With the online adaptive training, it is possible to achieve high-quality brain MR images which can represent the personalized variations. Although the method is new and experiments show promising results, there are limitations that are not covered in this paper.

Firstly, we assume that the disease progression is linear in the latent space the model learned. However, it might not be enough to fully model the progression of the brain in the real world. More sophisticated non-linear modeling of the brain progression with a larger dataset will be interesting future work.

Second, the assessment of the predicted images is limited to the evaluation with image quality metrics. In other words, we use SSIM and MSE metrics for evaluating the proposed method. It is reasonable to measure the structural similarity between the actual image and the predicted image. But it will be a meaningful extension to explore the user evaluation and follow-up use cases.

Third, the amount of training data is limited. Although we use the largest public longitudinal dataset (i.e., ADNI), the number of MR images is 426 subjects in this study. For that reason, we design this study based on 2D slice images instead of full 3D MR images. In addition, the limited resolution of brain image with 64 × 64 is used in this study. Further study with the larger dataset to extend the idea of personalized memory to high-resolution 3D MRI prediction will be the meaningful research direction.

Nevertheless, this paper shows the new personalized memory-based longitudinal modeling of brain images and shows the effectiveness of the proposed method. To the best of our knowledge, this is the first study to use online adaptive training for personalized modeling in deep learning-based longitudinal MRI modeling. Moreover, the proposed method is not limited to capturing the lateral ventricle growth in brain images. It can be extended to predict any temporal changes such as other diseases, tumors, and other parts of the image. Extension of this method to other applications such as longitudinal image analysis [7], [11] and video analysis [5], [12], [13] will be interesting future work.

VI. CONCLUSION

In this work, we proposed a new method for predicting longitudinal brain MR images using personalized memory. Our method effectively preserved brain structure and encoded temporal changes of the brain in MR images, which enabled the model to predict future and missing scans in a better way. Moreover, the online adaptive training to model the personalized memory progression significantly boosted the quality of complex brain MR images and changes.

ACKNOWLEDGMENT

(Seong Tae Kim and Umut KüçüKaslan are co-first authors.) The Alzheimer's Disease Neuroimaging Initiative: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

REFERENCES

[1] *Adni1:complete 2yr 1.5t Standardized Dataset*, Alzheimer's Disease Neuroimaging Initiative. [Online]. Available: http://adni.loni.ucla.edu

[2] M. Bland, *An Introduction to Medical Statistics*. London, U.K.: Oxford Univ. Press, 2015.

[3] B. Borroni, M. Di Luca, and A. Padovani, "Predicting Alzheimer dementia in mild cognitive impairment patients," *Eur. J. Pharmacol.*, vol. 545, no. 1, pp. 73–80, Sep. 2006.

[4] C. Bowles, R. Gunn, A. Hammers, and D. Rueckert, "Modelling the progression of alzheimer's disease in MRI using generative adversarial networks," *Proc. SPIE*, vol. 10574, Mar. 2018, Art. no. 105741K.

[5] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab, "OperA: Attention-regularized transformers for surgical phase recognition," 2021, *arXiv:2103.03873*. [Online]. Available: http://arxiv.org/abs/2103.03873

[6] M. J. D. Leon, S. DeSanti, R. Zinkowski, P. D. Mehta, D. Pratico, S. Segal, C. Clark, D. Kerkman, J. DeBernardis, J. Li, and L. Lair, "MRI and CSF studies in the early diagnosis of Alzheimer's disease," *J. Internal Med.*, vol. 256, no. 3, pp. 205–223, 2004.

[7] S. Denner, A. Khakzar, M. Sajid, M. Saleh, Z. Spiclin, S. T. Kim, and N. Navab, "Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries*. Cham, Switzerland: Springer, 2021, pp. 111–121.

[8] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–32.

[9] Y. Du, S. Li, and I. Mordatch, "Compositional visual generation with energy based models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6637–6647.

[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] S. T. Kim, L. Goli, M. Paschali, A. Khakzar, M. Keicher, T. Czempiel, E. Burian, R. Braren, N. Navab, and T. Wendler, "Longitudinal quantitative assessment of COVID-19 infection progression from chest CTs," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, Mar. 2021, pp. 2–11.

[12] S. T. Kim and Y. M. Ro, "Facial dynamics interpreter network: What are the important relations between local dynamics for facial trait estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 464–480.

[13] S. T. Kim and Y. M. Ro, "Attended relation feature representation of facial dynamics for facial authentication," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1768–1778, Jul. 2019.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.

[15] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 10215–10224.

[16] T. Klymenko, S. T. Kim, K. Lauber, C. Kurz, G. Landry, N. Navab, and S. Albarqouni, "Butterfly-Net: Spatial-temporal architecture for medical image segmentation," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 616–620.

[17] J.-H. Lee, S. T. Kim, H. Lee, and Y. M. Ro, "Feature2Mass: Visual feature processing in latent space for realistic labeled mass generation," in *Proc. Eur. Conf. Comput. Vis. Workshop*, Sep. 2018, pp. 1–9.

[18] R. S. Liu, L. Lemieux, G. S. Bell, S. M. Sisodiya, S. D. Shorvon, J. W. Sander, and J. S. Duncan, "A longitudinal study of brain morphometrics using quantitative magnetic resonance imaging and difference image analysis," *NeuroImage*, vol. 20, no. 1, pp. 22–33, Sep. 2003.

[19] M. Louis, R. Couronne, I. Koval, B. Charlier, and S. Durrleman, "Riemannian geometry learning for disease progression modelling," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Springer, 2019, pp. 542–553.

[20] S. M. Nestor, R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J. L. Wells, J. Fogarty, and R. Bartha, "Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database," *Brain*, vol. 131, no. 9, pp. 2443–2454, Aug. 2008.

[21] S. Pathan and Y. Hong, "Predictive image regression for longitudinal studies with missing data," in *Proc. Med. Imag. Deep Learn. (MIDL)*, Amsterdam, The Netherlands, 2018.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 97–108.

[23] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*. [Online]. Available: http://arxiv.org/abs/1710.05941

[24] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis," *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[25] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1530–1538.

[26] A. Sankar, M. Keicher, R. Eisawy, A. Parida, F. Pfister, S. T. Kim, and N. Navab, "GLOWin: A flow-based invertible generative framework for learning disentangled feature representations in medical images," 2021, *arXiv:2103.10868*. [Online]. Available: http://arxiv.org/abs/2103.10868

[27] M. Tanaka, "Weighted sigmoid gate unit for an activation function of deep neural network," *Pattern Recognit. Lett.*, vol. 135, pp. 354–359, Jul. 2020.

[28] H.-J. Tien, H.-C. Yang, P.-W. Shueng, and J.-C. Chen, "Cone-beam CT image quality improvement using cycle-deblur consistent adversarial networks (cycle-deblur GAN) for chest CT imaging in breast cancer patients," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Dec. 2021.

[29] A. R. Venkatakrishnan, S. T. Kim, R. Eisawy, F. Pfister, and N. Navab, "Self-supervised out-of-distribution detection in brain CT scans," 2020, *arXiv:2011.05428*. [Online]. Available: http://arxiv.org/abs/2011.05428

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

SEONG TAE KIM (Member, IEEE) received the B.S. degree from Korea University, Seoul, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2014 and 2019, respectively. In 2015, he was a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. From 2019 to 2021, he was a Senior Research Scientist with the Chair for Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, South Korea. He authored or coauthored more than 40 peer-reviewed journal articles and conference papers. His current research interests include deep learning, spatio-temporal learning, explainable deep learning, and medical image analysis. He received the Best Student Paper Award of SPIE Medical Imaging in 2018.

UMUT KÜÇÜKASLAN received the B.S. degree from Boğaziçi University, İstanbul, Turkey, and the M.S. degree from Technische Universität München, Munich, Germany. His thesis, supervised by Prof. Seong Tae Kim, was on the topic of "Deep Generative Models for Longitudinal Analysis of Alzheimer's Disease." He is currently a Machine Learning Engineer with Logivations GmbH, Munich. His current interests include deep learning, object detection and segmentation in warehouses, and marker tracking for real-time applications.

NASSIR NAVAB (Senior Member, IEEE) received the Ph.D. degree from INRIA and the University of Paris XI, France. He held a postdoctoral fellowship position at the MIT Media Laboratory, before joining Siemens Corporate Research (SCR), in 1994. He is currently a Full Professor and the Director of the Laboratory for Computer-Aided Medical Procedures, Johns Hopkins University, and the Technical University of Munich. He has also secondary faculty appointments at both affiliated Medical Schools. He is the author of hundreds of peer-reviewed scientific articles, with more than 45,000 citations and an H-index of 96 as of October 2021. He is the author of more than 30 awarded articles, including 11 at MICCAI, five at IPCAI, and three at IEEE ISMAR. He is the inventor of 50 granted U.S. patents and more than 50 international ones. His current research interests include medical augmented reality, computer-aided surgery, medical robotics, and machine learning. At SCR, he was a Distinguished Member and received the Siemens Inventor of the Year Award in 2001. He received the SMIT Society Technology Award in 2010 for the Introduction of Camera Augmented Mobile C-arm and Freehand SPECT technologies, the '10 years Lasting Impact Award' of IEEE ISMAR in 2015, and MICCAI Enduring Impact Award in 2021. In 2012, he was elected as a fellow of the MICCAI Society. He has acted as a member of the board of directors of the MICCAI Society, 2007–2012 and 2014–2017, and serves on the Steering Committee of the IEEE Symposium on Mixed and Augmented Reality (ISMAR) and Information Processing in Computer-Assisted Interventions (IPCAI).