This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

Combining Noise-to-Image and Image-to-Image GANs: Brain MR Image Augmentation for Tumor Detection

CHANGHEE HAN[1,2,3], LEONARDO RUNDO[3,4,5], RYOSUKE ARAKI[6], YUDAI NAGANO[1], YUJIRO FURUKAWA[7], GIANCARLO MAURI[5], HIDEKI NAKAYAMA[1,8], HIDEAKI HAYASHI[2,9]

[1]Machine Perception Group, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8657, Japan
[2]Research Center for Medical Big Data, National Institute of Informatics, Tokyo 100-0003, Japan
[3]Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, United Kingdom
[4]Cancer Research UK Cambridge Institute, Cambridge CB2 0RE, United Kingdom
[5]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan 20126, Italy
[6]Machine Perception and Robotics Group, Graduate School of Engineering, Chubu University, Aichi 487-8501, Japan
[7]Department of Psychiatry, Jikei University School of Medicine, Tokyo 105-8461, Japan
[8]International Research Center for Neurointelligence (WPI-IRCN), Institutes for Advanced Study, The University of Tokyo, Tokyo 113-8657, Japan
[9]Human Interface Laboratory, Department of Advanced Information Technology, Kyushu University, Fukuoka 819-0395, Japan

Corresponding author: Changhee Han (e-mail: han@nlab.ci.i.u-tokyo.ac.jp).

ABSTRACT Convolutional Neural Networks (CNNs) achieve excellent computer-assisted diagnosis with sufficient annotated training data. However, most medical imaging datasets are small and fragmented. In this context, Generative Adversarial Networks (GANs) can synthesize realistic/diverse additional training images to fill the data lack in the real image distribution; researchers have improved classification by augmenting data with noise-to-image (e.g., random noise samples to diverse pathological images) or image-to-image GANs (e.g., a benign image to a malignant one). Yet, no research has reported results combining noise-to-image and image-to-image GANs for further performance boost. Therefore, to maximize the DA effect with the GAN combinations, we propose a two-step GAN-based DA that generates and refines brain Magnetic Resonance (MR) images with/without tumors separately: (*i*) Progressive Growing of GANs (PGGANs), multi-stage noise-to-image GAN for high-resolution MR image generation, first generates realistic/diverse 256×256 images; (*ii*) Multimodal UNsupervised Image-to-image Translation (MUNIT) that combines GANs/Variational AutoEncoders or SimGAN that uses a DA-focused GAN loss, further refines the texture/shape of the PGGAN-generated images similarly to the real ones. We thoroughly investigate CNN-based tumor classification results, also considering the influence of pre-training on ImageNet and discarding weird-looking GAN-generated images. The results show that, when combined with classic DA, our two-step GAN-based DA can significantly outperform the classic DA alone, in tumor detection (i.e., boosting sensitivity 93.67% to 97.48%) and also in other medical imaging tasks.

INDEX TERMS Data augmentation, Synthetic image generation, GANs, Brain MRI, Tumor detection

I. INTRODUCTION

Convolutional Neural Networks (CNNs) are playing a key role in medical image analysis, updating the state-of-the-art in many tasks [1]–[3] when large-scale annotated training data are available. However, preparing such massive medical data is demanding; thus, for better diagnosis, researchers generally adopt classic Data Augmentation (DA) techniques, such as geometric/intensity transformations of original images [4], [5]. Those augmented images, however, intrinsically have a similar distribution to the original ones, resulting in limited performance improvement. In this sense, Generative Adversarial Network (GAN)-based DA can considerably increase the performance [6]; since the generated images are realistic but completely new samples, they can fill the real image distribution uncovered by the original dataset [7].

The main problem in computer-assisted diagnosis lies in small/fragmented medical imaging datasets from multiple scanners; thus, researchers have improved classification by augmenting images with noise-to-image GANs (e.g., random noise samples to diverse pathological images [8]) or image-to-image GANs (e.g., a benign image to a malignant one [9]). However, no research has achieved further performance boost by combining noise-to-image and image-to-image GANs.

So, how can we maximize the DA effect under limited training images using the GAN combinations? To generate and refine brain Magnetic Resonance (MR) images with/without tumors separately, we propose a two-step GAN-based DA approach: (*i*) Progressive Growing of GANs (PGGANs) [10], low-to-high resolution noise-to-image GAN, first generates realistic/diverse 256×256 images—the PG-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

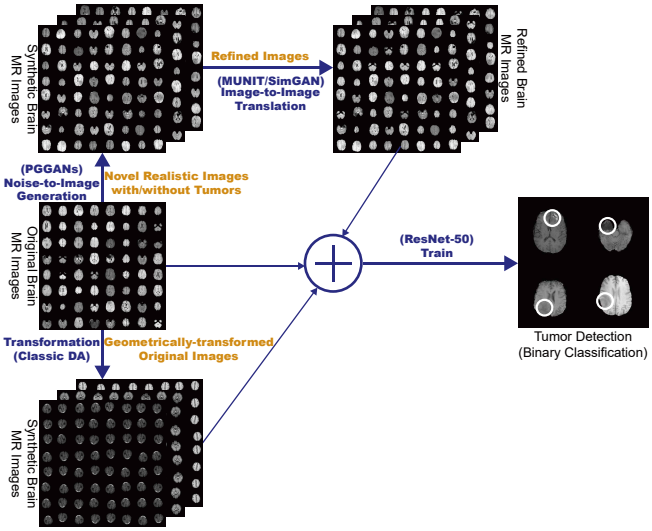

FIGURE 1: Combining noise-to-image and image-to-image GAN for better tumor detection: the PGGANs generates a number of realistic brain tumor/non-tumor MR images separately, the MUNIT/SimGAN refines them separately, and the binary classifier uses them as additional training data.

GANs helps DA since most CNN architectures adopt around 256×256 input sizes (e.g., InceptionResNetV2 [11]: 299×299, ResNet-50 [12]: 224×224); (*ii*) Multimodal UNsupervised Image-to-image Translation (MUNIT) [13] that combines GANs/Variational AutoEncoders (VAEs) [14] or SimGAN [15] that uses a DA-focused GAN loss, further refines the texture/shape of the PGGAN-generated images to fit them into the real image distribution. Since training a single sophisticated GAN system is already difficult, instead of end-to-end training, we adopt a two-step approach for performance boost *via* an ensemble generation process from those state-of-the-art GANs' different algorithms.

We thoroughly investigate CNN-based tumor classification results, also considering the influence of pre-training on ImageNet [16] and discarding weird-looking GAN-generated images. Moreover, we evaluate the synthetic images' appearance *via* Visual Turing Test [17] by an expert physician, and visualize the data distribution of real/synthetic images *via* t-Distributed Stochastic Neighbor Embedding (t-SNE) [18]. When combined with classic DA, our two-step GAN-based DA approach significantly outperforms the classic DA alone, boosting sensitivity 93.67% to 97.48%[1].

Research Questions. We mainly address two questions:

- **GAN Selection:** Which GAN architectures are well-suited for realistic/diverse medical image generation?
- **Medical DA:** How to use GAN-generated images as additional training data for better CNN-based diagnosis?

[1]This paper remarkably improves our preliminary work [8] investigating the potential of the ImageNet-pre-trained PGGANs—with minimal preprocessing and no refinement—for DA using a vanilla version of ResNet-50, resulting in minimum performance boost; since PGGAN-generated images unstabilized ResNet-50 training, we further optimize the ResNet-50 hyperparameters (i.e., the optimizer, learning rate, and decay rate) according to the training data, also modifying its architecture before the final sigmoid layer.

Contributions. Our main contributions are as follows:

- **Whole Image Generation:** This research shows that PGGANs can generate realistic/diverse 256×256 whole medical images—not only small pathological sub-areas—and MUNIT can further refine their texture/shape similarly to real ones.
- **Two-step GAN-based DA:** This novel two-step approach, combining for the first time noise-to-image and image-to-image GANs, significantly boosts tumor detection sensitivity.
- **Misdiagnosis Prevention:** This study firstly analyzes how medical GAN-based DA is associated with pre-training on ImageNet and discarding weird-looking synthetic images to achieve high sensitivity with small/fragmented datasets.

The manuscript is organized as follows. Section II covers the background of GANs, especially focusing on GAN-based DA in medical imaging. Section III describes the analyzed brain tumor MRI dataset, along with the investigated image generation method using a noise-to-image GAN (i.e., PGGANs) and refinement methods using image-to-image GANs (i.e., MUNIT and SimGAN), respectively. This section also explains how to evaluate those synthesized images based on tumor detection *via* ResNet-50, clinical validation *via* Visual Turing Test, and visualization *via* t-SNE. Section IV presents and discusses the experimental results. Lastly, Section V provides the conclusive remarks and future directions.

II. GENERATIVE ADVERSARIAL NETWORKS

VAEs [14] often accompany blurred samples despite easier training, due to the imperfect reconstruction using a single objective function; meanwhile, GANs [6] have revolutionized image generation in terms of realism/diversity [19] based on a two-player objective function: a generator G tries to generate realistic images to fool a discriminator D while maintaining diversity; D attempts to distinguish between the real/synthetic images. However, difficult GAN training from the two-player objective function accompanies artifacts/mode collapse [20], when generating high-resolution images (e.g., 256×256 pixels) [21]; to tackle this, multi-stage noise-to-image GANs have been proposed: AttnGAN [22] generates images from text using attention-based multi-stage refinement; PGGANs [10] generates realistic images using low-to-high resolution multi-stage training. Contrarily, to obtain images with desired texture/shape, researchers have proposed image-to-image GANs: MUNIT [13] translates images using both GANs/VAEs; SimGAN [15] translates images for DA using the self-regularization term/local adversarial loss.

Especially in medical imaging, to handle small and fragmented datasets from multiple scanners, researchers have exploited both noise-to-image and image-to-image GANs as DA techniques to improve classification: researchers used the noise-to-image GANs to augment liver lesion Computed Tomography (CT) [23] and chest cardiovascular abnormality X-ray images [24]; others used the image-to-image GANs to augment breast cancer mammography images [9] and bone

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

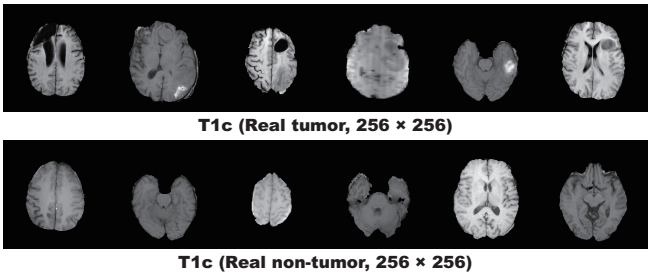

FIGURE 2: Example real MR images used for PGGAN training.

lesion X-ray images [25], translating benign images into malignant ones and *vice versa*.

However, to the best of our knowledge, we are the first to combine noise-to-image and image-to-image GANs to maximize the DA performance. Moreover, this is the first medical GAN work generating whole 256×256 images, instead of regions of interest (i.e., small pathological sub-areas) alone, for robust classification. Along with classic image transformations, a novel approach—augmenting realistic/diverse whole medical images with the two-step GAN—may become a clinical breakthrough.

III. MATERIALS AND METHODS

A. BRATS 2016 TRAINING SET

We use a dataset of 240×240 contrast-enhanced T1-weighted (T1c) brain axial MR images of 220 high-grade glioma cases from the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) 2016 [26]. T1c is the most common sequence in tumor detection thanks to its high-contrast [27].

B. PGGAN-BASED IMAGE GENERATION

Pre-processing For better GAN/ResNet-50 training, we select the slices from #30 to #130 among the whole 155 slices to omit initial/final slices, which convey negligible useful information; also, since tumor/non-tumor annotation in the BRATS 2016 dataset, based on 3D volumes, is highly incorrect/ambiguous on 2D slices, we exclude (i) tumor images tagged as non-tumor, (ii) non-tumor images tagged as tumor, (iii) borderline images with unclear tumor/non-tumor appearance, and (iv) images with missing brain parts due to the skull-stripping procedure[2]. For tumor detection, we divide the whole dataset (220 patients) into:

- Training set
  (154 patients/4,679 tumor/3,750 non-tumor images);
- Validation set
  (44 patients/750 tumor/608 non-tumor images);
- Test set
  (22 patients/1,232 tumor/1,013 non-tumor images).

During the GAN training, we only use the training set to be fair; for better PGGAN training, the training set images are zero-padded to reach a power of 2: 256×256 pixels from 240×240. Fig. 2 shows example real MR images.

[2]Although this discarding procedure could be automated, we manually conduct it for reliability; the pre-processed dataset is available on Dropbox in https://www.dropbox.com/s/v208w1q7kpvv9lo/data.zip?dl=0

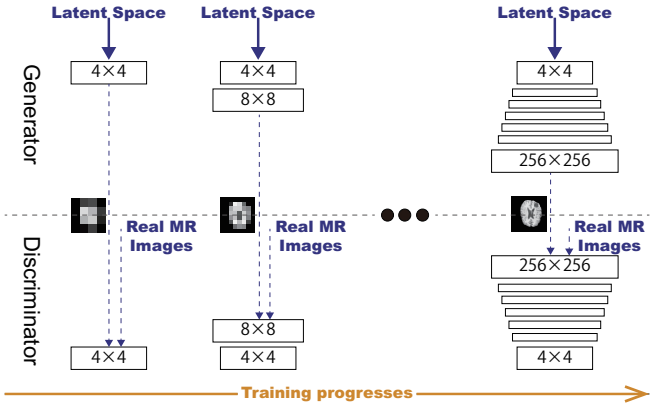

FIGURE 3: PGGAN architecture for 256×256 brain MR image generation. $N \times N$ refers to convolutional layers operating on $N \times N$ spatial resolution.

PGGANs [10] is a GAN training method that progressively grows a generator and discriminator: starting from low resolution, new layers model details as training progresses. This study adopts the PGGANs to synthesize realistic/diverse 256×256 brain MR images (Fig. 3); we train and generate tumor/non-tumor images separately.

PGGAN Implementation Details The PGGAN architecture adopts the Wasserstein loss with gradient penalty [20]:

$$\mathop{\mathbb{E}}_{\tilde{\mathbf{y}} \sim \mathbb{P}_g}[D(\tilde{\mathbf{y}})] - \mathop{\mathbb{E}}_{\mathbf{y} \sim \mathbb{P}_r}[D(\mathbf{y})] + \lambda_{\text{gp}} \mathop{\mathbb{E}}_{\hat{\mathbf{y}} \sim \mathbb{P}_{\hat{\mathbf{y}}}}[(\|\nabla_{\hat{\mathbf{y}}} D(\hat{\mathbf{y}})\|_2 - 1)^2], \quad (1)$$

where $\mathbb{E}[\cdot]$ denotes the expected value, the discriminator $D \in \mathcal{D}$ (i.e., the set of 1-Lipschitz functions), $\mathbb{P}_r$ is the data distribution defined by the true data sample $\mathbf{y}$, and $\mathbb{P}_g$ is the model distribution defined by the generated sample $\tilde{\mathbf{y}} = G(\mathbf{z})$ ($\mathbf{z} \sim p(\mathbf{z})$ is the input noise $\mathbf{z}$ to the generator sampled from a Gaussian distribution). A gradient penalty is added for the random sample $\hat{\mathbf{y}} \sim \mathbb{P}_{\hat{\mathbf{y}}}$, where $\nabla_{\hat{\mathbf{y}}}$ is the gradient operator towards the generated samples and λ_{gp} is the gradient penalty coefficient.

We train the model (Table 1) for 100 epochs with a batch size of 16 and 1.0×10^{-3} learning rate for the Adam optimizer (the exponential decay rates $\beta_1 = 0, \beta_2 = 0.99$) [29]. All experiments use $\lambda_{\text{gp}} = 10$ with 1 critic iteration per generator iteration. During training, we apply random cropping in 0-15 pixels as DA.

C. MUNIT/SIMGAN-BASED IMAGE REFINEMENT

Refinement Using resized 224×224 images for ResNet-50, we further refine the texture/shape of PGGAN-generated tumor/non-tumor images separately to fit them into the real image distribution using MUNIT [13] or SimGAN [15]. SimGAN remarkably improved eye gaze estimation results after refining non-GAN-based synthetic images from the UnityEyes simulator *via* image-to-image translation; thus, we also expect such performance improvement after refining synthetic images from a noise-to-image GAN (i.e., PGGANs) *via* an image-to-image GAN (i.e., MUNIT/SimGAN) with considerably different GAN algorithms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

TABLE 1: PGGAN architecture details for the generator/discriminator. Pixelwise feature vector normalization [28] is applied in the generator after each convolutional layer except for the final output layer as in the original paper [10]. LReLU denotes Leaky ReLU with leakiness 0.2.

Generator	Activation	Output Shape
Latent vector	–	512 × 1 × 1
Conv 4 × 4	LReLU	512 × 4 × 4
Conv 3 × 3	LReLU	512 × 4 × 4
Upsample	–	512 × 8 × 8
Conv 3 × 3	LReLU	512 × 8 × 8
Conv 3 × 3	LReLU	512 × 8 × 8
Upsample	–	512 × 16 × 16
Conv 3 × 3	LReLU	256 × 16 × 16
Conv 3 × 3	LReLU	256 × 16 × 16
Upsample	–	256 × 32 × 32
Conv 3 × 3	LReLU	128 × 32 × 32
Conv 3 × 3	LReLU	128 × 32 × 32
Upsample	–	128 × 64 × 64
Conv 3 × 3	LReLU	64 × 64 × 64
Conv 3 × 3	LReLU	64 × 64 × 64
Upsample	–	64 × 128 × 128
Conv 3 × 3	LReLU	32 × 128 × 128
Conv 3 × 3	LReLU	32 × 128 × 128
Upsample	–	32 × 256 × 256
Conv 3 × 3	LReLU	16 × 256 × 256
Conv 3 × 3	LReLU	16 × 256 × 256
Conv 1 × 1	Linear	1 × 256 × 256

Discriminator	Activation	Output Shape
Input image	–	1 × 256 × 256
Conv 1 × 1	LReLU	16 × 256 × 256
Conv 3 × 3	LReLU	16 × 256 × 256
Conv 3 × 3	LReLU	32 × 256 × 256
Downsample	–	32 × 128 × 128
Conv 3 × 3	LReLU	32 × 128 × 128
Conv 3 × 3	LReLU	64 × 128 × 128
Downsample	–	64 × 64 × 64
Conv 3 × 3	LReLU	64 × 64 × 64
Conv 3 × 3	LReLU	128 × 64 × 64
Downsample	–	128 × 32 × 32
Conv 3 × 3	LReLU	128 × 32 × 32
Conv 3 × 3	LReLU	256 × 32 × 32
Downsample	–	256 × 16 × 16
Conv 3 × 3	LReLU	256 × 16 × 16
Conv 3 × 3	LReLU	512 × 16 × 16
Downsample	–	512 × 8 × 8
Conv 3 × 3	LReLU	512 × 8 × 8
Conv 3 × 3	LReLU	512 × 8 × 8
Downsample	–	512 × 4 × 4
Minibatch stddev	–	513 × 4 × 4
Conv 3 × 3	LReLU	512 × 4 × 4
Conv 4 × 4	LReLU	512 × 1 × 1
Fully-connected	Linear	1 × 1 × 1

We randomly select $3,000$ real/$3,000$ PGGAN-generated tumor images for tumor image training, and we perform the same for non-tumor image training. To find suitable refining steps for each architecture, we pick the MUNIT/SimGAN models with the highest accuracy on tumor detection validation, when pre-trained and combined with classic DA, among $20,000$/$50,000$/$100,000$ steps, respectively.

MUNIT [13] is an image-to-image GAN based on both auto-encoding/translation; it extends UNIT [30] to increase the generated images' realism/diversity *via* a stochastic model representing continuous output distributions.

MUNIT Implementation Details The MUNIT architecture adopts the following loss:

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \quad \mathcal{L}_{\text{VAE}_1} + \mathcal{L}_{\text{GAN}_1} + \mathcal{L}_{\text{CC}_1} + \mathcal{L}_{\text{VGG}_1} + \mathcal{L}_{\text{VAE}_2} + \mathcal{L}_{\text{GAN}_2} + \mathcal{L}_{\text{CC}_2} + \mathcal{L}_{\text{VGG}_2}, \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes the loss function. Using the multiple encoders E_1/E_2, generators G_1/G_2, discriminators D_1/D_2, cycle-consistencies CC_1/CC_2, and domain-invariant perceptions VGG_1/VGG_2 [31], this framework jointly solves learning problems of the VAE_1/VAE_2 and GAN_1/GAN_2 for the image reconstruction streams, image translation streams, cycle-consistency reconstruction streams, and domain-invariant perception streams. Since we do not need the style loss for our experiments, instead of the MUNIT loss, we use the UNIT loss with the perceptual loss for the MUNIT architecture (as in the UNIT authors' GitHub repository).

We train the model (Table 2) for $100,000$ steps with a batch size of 1 and 1.0×10^{-4} learning rate for the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) [29]. The learning rate is reduced by half every $20,000$ steps. We use the following MUNIT weights: the adversarial loss weight $= 1$; the image reconstruction loss weight $= 10$; the Kullback-Leibler (KL) divergence loss weight for reconstruction $= 0.01$; the cycle consistency loss weight $= 10$; the KL divergence loss weight for cycle consistency $= 0.01$; the domain-invariant perceptual loss weight $= 1$; the Least Squares GAN objective function for the discriminators [34]. During training, we apply horizontal flipping as DA.

SimGAN [15] is an image-to-image GAN designed for DA that adopts the self-regularization term/local adversarial loss; it updates a discriminator with a history of refined images.

SimGAN Implementation Details The SimGAN architecture (i.e., a refiner) uses the following loss:

$$\sum_i \mathcal{L}_{\text{real}}(\boldsymbol{\theta}; \mathbf{x}_i, \mathcal{Y}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathbf{x}_i), \tag{3}$$

where $\mathcal{L}(\cdot)$ denotes the loss function, $\boldsymbol{\theta}$ is the function parameters, $\mathbf{x}_i$ is the i^{th} PGGAN-generated training image, and $\mathcal{Y}$ is the set of the real images $\mathbf{y}_j$. The first part $\mathcal{L}_{\text{real}}$ adds realism to the synthetic images using a discriminator, while the second part $\mathcal{L}_{\text{reg}}$ preserves the tumor/non-tumor features.

We train the model (Table 3) for $20,000$ steps with a batch size of 10 and 1.0×10^{-4} learning rate for the Stochastic Gradient Descent (SGD) optimizer [35] without momentum.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

TABLE 2: MUNIT architecture details for the generator/discriminator. We input color images (i.e., 3 channels) to use ImageNet initialization. Instance normalization [32]/adaptive instance normalization [33] are applied in the content encoder/decoder after each convolutional layer respectively except for the final decoder output layer as in the original paper [13]. LReLU denotes Leaky ReLU with leakiness 0.2.

Generator	**Activation**	**Output Shape**
Content Encoder		
Input image	–	3 × 224 × 224
Conv 7 × 7	ReLU	64 × 224 × 224
Conv 4 × 4	ReLU	128 × 112 × 112
Conv 4 × 4	ReLU	256 × 56 × 56
ResBlock $\begin{bmatrix} 3\times3 \\ 3\times3 \end{bmatrix}\times4$	ReLU	256 × 56 × 56
	–	256 × 56 × 56
Decoder		
ResBlock $\begin{bmatrix} 3\times3 \\ 3\times3 \end{bmatrix}\times4$	ReLU	256 × 56 × 56
	–	256 × 56 × 56
Upsample	–	256 × 112 × 112
Conv 5 × 5	ReLU	128 × 112 × 112
Upsample	–	128 × 224 × 224
Conv 5 × 5	ReLU	64 × 224 × 224
Conv 7 × 7	Tanh	3 × 224 × 224

Discriminator	**Activation**	**Output Shape**
Input image	–	3 × 224 × 224
Conv 4 × 4	LReLU	64 × 112 × 112
Conv 4 × 4	LReLU	128 × 56 × 56
Conv 4 × 4	LReLU	256 × 28 × 28
Conv 4 × 4	LReLU	512 × 14 × 14
Conv 4 × 4	–	1 × 14 × 14
AveragePool	–	3 × 112 × 112
Conv 4 × 4	LReLU	64 × 56 × 56
Conv 4 × 4	LReLU	128 × 28 × 28
Conv 4 × 4	LReLU	256 × 14 × 14
Conv 4 × 4	LReLU	512 × 7 × 7
Conv 4 × 4	–	1 × 7 × 7
AveragePool	–	3 × 56 × 56
Conv 4 × 4	LReLU	64 × 28 × 28
Conv 4 × 4	LReLU	128 × 14 × 14
Conv 4 × 4	LReLU	256 × 7 × 7
Conv 4 × 4	LReLU	512 × 3 × 3
Conv 4 × 4	–	1 × 3 × 3
AveragePool	–	3 × 28 × 28

TABLE 3: SimGAN architecture details for the refiner/discriminator. Batch normalization is applied both in the refiner/discriminator after each convolutional layer except for the final output layers respectively as in the original paper [15].

Refiner	**Activation**	**Output Shape**
Input image	–	1 × 224 × 224
Conv 9 × 9	ReLU	64 × 224 × 224
ResBlock $\begin{bmatrix} 3\times3 \\ 3\times3 \end{bmatrix}\times12$	ReLU	64 × 224 × 224
	–	64 × 224 × 224
Conv 1 × 1	Tanh	1 × 224 × 224

Discriminator	**Activation**	**Output Shape**
Input image	–	1 × 224 × 224
Conv 9 × 9	ReLU	96 × 72 × 72
Conv 5 × 5	ReLU	64 × 68 × 68
Maxpool	–	64 × 34 × 34
Conv 5 × 5	ReLU	64 × 15 × 15
Conv 3 × 3	ReLU	32 × 13 × 13
Maxpool	–	32 × 7 × 7
Conv 1 × 1	ReLU	32 × 7 × 7
Conv 1 × 1	ReLU	2 × 7 × 7

The learning rate is reduced by half at 15,000 steps. We train the refiner first with just the self-regularization loss with $\lambda_{\text{reg}} = 5 \times 10^{-5}$ for 500 steps; then, for each update of the discriminator, we update the refiner 5 times. During training, we apply horizontal flipping as DA.

D. TUMOR DETECTION USING RESNET-50

Pre-processing. As ResNet-50's input size is 224 × 224 pixels, we resize the whole real images from 240 × 240 and whole PGGAN-generated images from 256 × 256.

ResNet-50 [12] is a 50-layer residual learning-based CNN. We adopt it to detect brain tumors in MR images (i.e., the binary classification of tumor/non-tumpor images) due to its outstanding performance in image classification tasks [36], including binary classification [37]. Chang *et al.* [38] also used a similar 34-layer residual convolutional network for the binary classification of brain tumors (i.e., determining the Isocitrate Dehydrogenase status in low-/high-grade gliomas).

DA Setups To confirm the effect of PGGAN-based DA and its refinement using MUNIT/SimGAN, we compare the following 10 DA setups under sufficient images both with/without ImageNet [16] pre-training (i.e., 20 DA setups):

1) 8429 real images;
2) + 200k classic DA;
3) + 400k classic DA;
4) + 200k PGGAN-based DA;
5) + 200k PGGAN-based DA w/o clustering/discarding;
6) + 200k classic DA & 200k PGGAN-based DA;
7) + 200k MUNIT-refined DA;
8) + 200k classic DA & 200k MUNIT-refined DA;
9) + 200k SimGAN-refined DA;
10) + 200k classic DA & 200k SimGAN-refined DA.

Due to the risk of overlooking the tumor diagnosis *via* medical imaging, higher sensitivity matters much more than higher specificity [39]; thus, we aim to achieve higher sensitivity, using the additional synthetic training images. We perform McNemar's test on paired tumor detection results [40] to confirm our two-step GAN-based DA's statistically-significant sensitivity improvement; since this statistical analysis involves multiple comparison tests, we adjust their p-values using the Holm–Bonferroni method [41].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

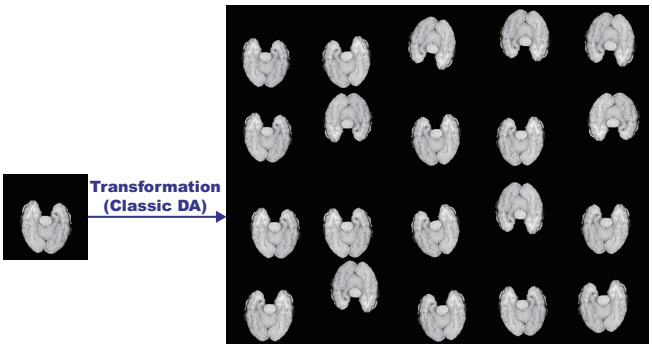

FIGURE 4: Example real MR image and its geometrically-transformed images.

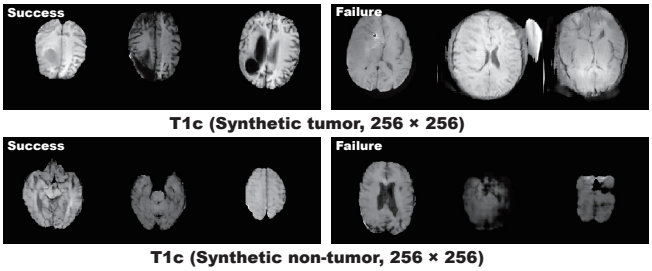

FIGURE 5: Example PGGAN-generated MR images: (a) Success cases; (b) Failure cases.

Whereas medical imaging researchers widely use the ImageNet initialization despite different textures of natural/medical images, recent study found that such ImageNet-trained CNNs are biased towards recognizing texture rather than shape [42]; thus, we aim to investigate how the medical GAN-based DA affects classification performance with/without the pre-training. As the classic DA, we adopt a random combination of horizontal/vertical flipping, rotation up to 10 degrees, width/height shift up to 8%, shearing up to 8%, zooming up to 8%, and constant filling of points outside the input boundaries (Fig. 4). For the PGGAN-based DA and its refinement, we only use success cases after discarding weird-looking synthetic images (Fig. 5); DenseNet-169 [43] extracts image features and k-means++ [44] clusters the features into 200 groups, and then we manually discard each cluster containing similar weird-looking images. To verify its effect, we also conduct a PGGAN-based DA experiment without the discarding step.

ResNet-50 Implementation Details The ResNet-50 architecture adopts the binary cross-entropy loss for binary classification both with/without ImageNet pre-training. As shown in Table 4, for robust training, before the final sigmoid layer, we introduce a 0.5 dropout [45], linear dense, and batch normalization [46] layers—training with GAN-based DA tends to be unstable especially without the batch normalization layer. We use a batch size of 96, 1.0×10^{-2} learning rate for the SGD optimizer [35] with 0.9 momentum, and early stopping of 20 epochs. The learning rate was multiplied by 0.1 every 20 epochs for the training from scratch and by 0.5 every 5 epochs for the ImageNet pre-training.

TABLE 4: ResNet-50 architecture details without/with pre-training. We input grayscale images (i.e., 1 channel) for experiments without pre-training, whereas we input color images (i.e., 3 channels) for experiments with pre-training to use ImageNet initialization. Batch normalization is applied after each convolutional layer as in the original paper [12].

Classifier			Activation	Output Shape
Input image			–	1 (3)× 224 × 224
Conv 7 × 7			ReLU	64 × 112 × 112
Maxpool			–	64 × 55 × 55
ResBlock	1×1 3×3 1×1	×3	ReLU ReLU ReLU	64 × 55 × 55 64 × 55 × 55 256 × 55 × 55
ResBlock	1×1 3×3 1×1	×4	ReLU ReLU ReLU	128 × 28 × 28 128 × 28 × 28 512 × 28 × 28
ResBlock	1×1 3×3 1×1	×6	ReLU ReLU ReLU	256 × 14 × 14 256 × 14 × 14 1024× 14 × 14
ResBlock	1×1 3×3 1×1	×3	ReLU ReLU ReLU	512 × 7 × 7 512 × 7 × 7 2048× 7 × 7
AveragePool			–	2048× 1 × 1
Flatten			–	2048
0.5 Dropout			–	2048
Dense			–	2
BatchNorm			Sigmoid	2

E. CLINICAL VALIDATION USING VISUAL TURING TEST

To quantify the (*i*) realism of 224×224 synthetic images by PGGANs, MUNIT, and SimGAN against real images respectively (i.e., 3 setups) and (*ii*) clearness of their tumor/non-tumor features, we supply, in random order, to an expert physician a random selection of:

- 50 real tumor images;
- 50 real non-tumor images;
- 50 synthetic tumor images;
- 50 synthetic non-tumor images.

Then, the physician has to classify them as both (*i*) real/synthetic and (*ii*) tumor/non-tumor, without previously knowing which is real/synthetic and tumor/non-tumor. The so-called Visual Turing Test [17] can probe the human ability to identify attributes and relationships in images, also for visually evaluating GAN-generated images [15]; this also applies to medical images for clinical decision-making tasks [47], [48], wherein physicians' expertise is critical.

F. VISUALIZATION USING T-SNE

To visualize distributions of geometrically-transformed and each GAN-based 224×224 images by PGGANs, MUNIT, and SimGAN against real images respectively (i.e., 4 setups), we adopt t-SNE [18] on a random selection of:

- 300 real tumor images;
- 300 real non-tumor images;
- 300 geometrically-transformed or each GAN-based tumor images;
- 300 geometrically-transformed or each GAN-based non-tumor images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

TABLE 5: ResNet-50 tumor detection (i.e., binary classification) results with various DA, with (without) ImageNet pre-training. Sensitivity and specificity consider the slight tumor/non-tumor class imbalance (about 6:5) in the test set. Boldface indicates the best performance.

DA Setups	Accuracy (%)	Sensitivity (%)	Specificity (%)
(1) 8,429 real images	93.14 (86.33)	90.91 (88.88)	95.85 (83.22)
(2) + 200k classic DA	95.01 (92.20)	93.67 (89.94)	96.64 (94.97)
(3) + 400k classic DA	94.83 (93.23)	91.88 (90.91)	98.42 (96.05)
(4) + 200k PGGAN-based DA	93.94 (86.19)	92.61 (87.26)	95.56 (84.90)
(5) + 200k PGGAN-based DA w/o clustering/discarding	94.83 (80.67)	91.88 (80.19)	98.42 (81.24)
(6) + 200k classic DA & 200k PGGAN-based DA	96.17 (95.59)	93.99 (94.16)	**98.82** (97.33)
(7) + 200k MUNIT-refined DA	94.30 (83.65)	93.02 (87.82)	95.85 (78.58)
(8) + 200k classic DA & 200k MUNIT-refined DA	**96.70** (96.35)	95.45 (**97.48**)	98.22 (94.97)
(9) + 200k SimGAN-refined DA	94.48 (77.64)	92.29 (82.31)	97.14 (71.96)
(10) + 200k classic DA & 200k SimGAN-refined DA	96.39 (95.01)	95.13 (95.05)	97.93 (94.97)

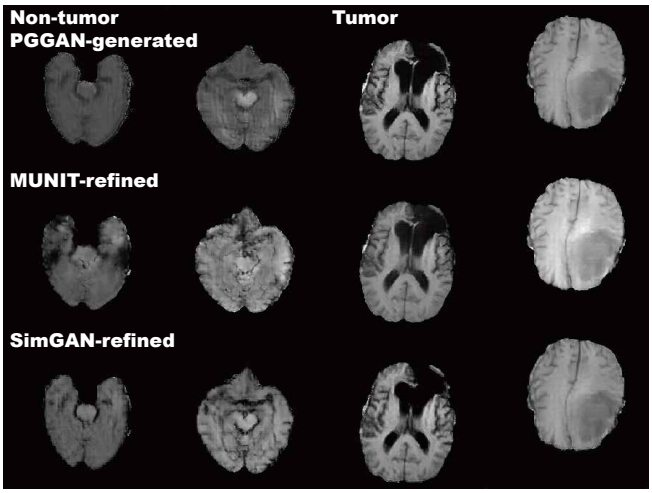

FIGURE 6: Example PGGAN-generated MR images and their refined versions by MUNIT/SimGAN.

We select only 300 images per each category for better visualization. The t-SNE method reduces the dimensionality to represent high-dimensional data into a lower-dimensional (2D/3D) space; it non-linearly balances between the input data's local and global aspects using perplexity.

T-SNE Implementation Details The t-SNE uses a perplexity of 100 for $1,000$ iterations to visually represent a 2D space. We input the images after normalizing pixel values to $[0, 1]$. For point locations of the real images, we compress all the images simultaneously and plot each setup (i.e., the geometrically-transformed or each GAN-based images against the real ones) separately; we maintain their locations by projecting all the data onto the same subspace.

IV. RESULTS

This section shows how PGGANs generates synthetic brain MR images and how MUNIT and SimGAN refine them. The results include instances of synthetic images, their quantitative evaluation by an expert physician, their t-SNE visualization, and their influence on tumor detection.

A. MR IMAGES GENERATED BY PGGANS

Fig. 5 illustrates examples of synthetic MR images by PGGANs. We visually confirm that, for about 75% of cases, it successfully captures the T1c-specific texture and tumor appearance, while maintaining the realism of the original brain MR images; but, for the rest 25%, the generated images lack clear tumor/non-tumor features or contain unrealistic features (i.e., hyper-intensity, gray contours, and odd artifacts).

B. MR IMAGES REFINED BY MUNIT/SIMGAN

MUNIT and SimGAN differently refine PGGAN-generated images—they render the texture and contours while maintaining the overall shape (Fig. 6). Non-tumor images change more remarkably than tumor images for both MUNIT and SimGAN; it probably derives from unsupervised image translation's loss for consistency to avoid image collapse, resulting in conservative change for more complicated images.

C. TUMOR DETECTION RESULTS

Table 5 shows the brain tumor classification results with/without DA while Table 6 indicates their pairwise comparison (p-values between our two-step GAN-based DA setups and the other DA setups) using McNemar's test. ImageNet pre-training generally outperforms training from scratch despite different image domains (i.e., natural images to medical images). As expected, classic DA remarkably improves classification, while no clear difference exists between the $200,000$/$400,000$ classic DA under sufficient geometrically-transformed training images. When pre-trained, each GAN-based DA (i.e., PGGANs/MUNIT/SimGAN) alone helps classification due to the robustness from GAN-generated images; but, without pre-training, it harms classification due to the biased initialization from the GAN-overwhelming data distribution. Similarly, without pre-training, PGGAN-based DA without clustering/discarding causes poor classification due to the synthetic images with severe artifacts, unlike the PGGAN-based DA's comparable results with/without the discarding step when pre-trained.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

TABLE 6: McNemar's test p-values for the pairwise comparison of the ResNet-50 tumor detection results in terms of accuracy, sensitivity, specificity, respectively. We compare our two-step GAN-based DA setups and all the other DA setups. All numbers within parentheses refer to DA setups on Table 5 and PT denotes pre-training. Boldface indicates statistical significance (threshold p-value < 0.05).

DA Setup Comparison	Accu	Sens	Spec	DA Setup Comparison	Accu	Sens	Spec	DA Setup Comparison	Accu	Sens	Spec
(7) w/ PT *vs* (1) w/ PT	0.693	0.206	1	(7) w/ PT *vs* (1) w/o PT	**< 0.001**	**0.002**	**< 0.001**	(7) w/ PT *vs* (2) w/ PT	1	1	1
(7) w/ PT *vs* (2) w/o PT	**0.034**	**0.024**	1	(7) w/ PT *vs* (3) w/ PT	1	1	**0.035**	(7) w/ PT *vs* (3) w/o PT	1	0.468	1
(7) w/ PT *vs* (4) w/ PT	1	1	1	(7) w/ PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/ PT *vs* (5) w/ PT	1	1	**0.003**
(7) w/ PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/ PT *vs* (6) w/ PT	**0.009**	1	**< 0.001**	(7) w/ PT *vs* (6) w/o PT	0.397	1	1
(7) w/ PT *vs* (7) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/ PT *vs* (8) w/ PT	**< 0.001**	**0.025**	**0.045**	(7) w/ PT *vs* (8) w/o PT	**0.008**	**< 0.001**	1
(7) w/ PT *vs* (9) w/ PT	1	1	1	(7) w/ PT *vs* (9) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/ PT *vs* (10) w/ PT	**< 0.001**	0.077	0.108
(7) w/ PT *vs* (10) w/o PT	1	0.206	1	(7) w/o PT *vs* (1) w/ PT	**< 0.001**	0.135	**< 0.001**	(7) w/o PT *vs* (1) w/o PT	**0.026**	1	**0.014**
(7) w/o PT *vs* (2) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/o PT *vs* (2) w/o PT	**< 0.001**	1	**< 0.001**	(7) w/o PT *vs* (3) w/ PT	**< 0.001**	**0.020**	**< 0.001**
(7) w/o PT *vs* (3) w/o PT	**< 0.001**	0.147	**< 0.001**	(7) w/o PT *vs* (4) w/ PT	**< 0.001**	**0.002**	**< 0.001**	(7) w/o PT *vs* (4) w/o PT	**0.044**	1	**< 0.001**
(7) w/o PT *vs* (5) w/ PT	**< 0.001**	**0.015**	**< 0.001**	(7) w/o PT *vs* (5) w/o PT	**0.011**	**< 0.001**	1	(7) w/o PT *vs* (6) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**
(7) w/o PT *vs* (6) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/o PT *vs* (8) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/o PT *vs* (8) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(7) w/o PT *vs* (9) w/ PT	**< 0.001**	**0.004**	**< 0.001**	(7) w/o PT *vs* (9) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(7) w/o PT *vs* (10) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**
(7) w/o PT *vs* (10) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/ PT *vs* (1) w PT	**< 0.001**	**< 0.001**	**0.010**	(8) w/ PT *vs* (1) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(8) w/ PT *vs* (2) w/ PT	**< 0.001**	0.074	0.206	(8) w/ PT *vs* (2) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/ PT *vs* (3) w/ PT	**0.002**	**< 0.001**	1
(8) w/ PT *vs* (3) w/o PT	**< 0.001**	**< 0.001**	0.112	(8) w/ PT *vs* (4) w/ PT	**< 0.001**	**< 0.001**	**0.006**	(8) w/ PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(8) w/ PT *vs* (5) w/ PT	**0.002**	**< 0.001**	1	(8) w/ PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/ PT *vs* (6) w/ PT	1	0.128	1
(8) w/ PT *vs* (6) w/o PT	0.222	0.760	1	(8) w/ PT *vs* (8) w/o PT	1	**0.008**	**< 0.001**	(8) w/ PT *vs* (9) w/ PT	**< 0.001**	**< 0.001**	1
(8) w/ PT *vs* (9) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/ PT *vs* (10) w/ PT	1	1	1	(8) w/ PT *vs* (10) w/o PT	**0.007**	1	0
(8) w/o PT *vs* (1) w/ PT	**< 0.001**	**< 0.001**	1	(8) w/o PT *vs* (1) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (2) w/ PT	0.179	**< 0.001**	0.588
(8) w/o PT *vs* (2) w/o PT	**< 0.001**	**< 0.001**	1	(8) w/o PT *vs* (3) w/ PT	0.101	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (3) w/o PT	**< 0.001**	**< 0.001**	1
(8) w/o PT *vs* (4) w/ PT	**< 0.001**	**< 0.001**	1	(8) w/o PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (5) w/ PT	0.197	**< 0.001**	**< 0.001**
(8) w/o PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (6) w/ PT	1	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (6) w/o PT	1	**< 0.001**	**0.007**
(8) w/o PT *vs* (9) w/ PT	**0.023**	**< 0.001**	0.256	(8) w/o PT *vs* (9) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(8) w/o PT *vs* (10) w/ PT	1	**0.002**	**< 0.001**
(8) w/o PT *vs* (10) w/o PT	0.143	**0.005**	1	(9) w/ PT *vs* (1) w/ PT	0.387	1	1	(9) w/ PT *vs* (1) w/o PT	**< 0.001**	**0.046**	**< 0.001**
(9) w/ PT *vs* (2) w/ PT	1	1	1	(9) w/ PT *vs* (2) w/o PT	**0.008**	0.262	0.321	(9) w/ PT *vs* (3) w/ PT	1	1	0.931
(9) w/ PT *vs* (3) w/o PT	0.910	1	1	(9) w/ PT *vs* (4) w/ PT	1	1	0.764	(9) w/ PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(9) w/ PT *vs* (5) w/ PT	1	1	0.639	(9) w/ PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/ PT *vs* (6) w/ PT	**0.014**	0.660	0.066
(9) w/ PT *vs* (6) w/o PT	0.716	0.365	1	(9) w/ PT *vs* (9) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/ PT *vs* (10) w/ PT	**0.004**	**0.006**	1
(9) w/ PT *vs* (10) w/o PT	1	**0.017**	0.256	(9) w/o PT *vs* (1) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (1) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(9) w/o PT *vs* (2) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (2) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (3) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**
(9) w/o PT *vs* (3) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (4) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(9) w/o PT *vs* (5) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (5) w/o PT	**0.022**	1	**< 0.001**	(9) w/o PT *vs* (6) w/ PT	**< 0.001**	**< 0.001**	**< 0.001**
(9) w/o PT *vs* (6) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (10) w/PT	**< 0.001**	**< 0.001**	**< 0.001**	(9) w/o PT *vs* (10) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(10) w/ PT *vs* (1) w/ PT	**< 0.001**	**< 0.001**	**0.049**	(10) w/ PT *vs* (1) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(10) w/ PT *vs* (2) w/ PT	**0.039**	0.515	1
(10) w/ PT *vs* (2) w/o PT	**< 0.001**	**< 0.001**	**0.002**	(10) w/ PT *vs* (3) w/ PT	**0.017**	**< 0.001**	1	(10) w/ PT *vs* (3) w/o PT	**< 0.001**	**< 0.001**	0.415
(10) w/ PT *vs* (4) w/ PT	**< 0.001**	**0.019**	**0.028**	(10) w/ PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(10) w/ PT *vs* (5) w/ PT	**0.015**	**< 0.001**	1
(10) w/ PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(10) w/ PT *vs* (6) w/ PT	1	1	1	(10) w/ PT *vs* (6) w/o PT	0.981	1	1
(10) w/ PT *vs* (10) w/o PT	0.054	1	**0.002**	(10) w/o PT *vs* (1) w/ PT	**0.039**	**< 0.001**	1	(10) w/o PT *vs* (1) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(10) w/o PT *vs* (2) w/ PT	1	0.727	0.649	(10) w/o PT *vs* (2) w/o PT	**< 0.001**	**< 0.001**	1	(10) w/o PT *vs* (3) w/ PT	1	**0.002**	**< 0.001**
(10) w/o PT *vs* (3) w/o PT	**0.039**	**< 0.001**	1	(10) w/o PT *vs* (4) w/ PT	1	**0.019**	1	(10) w/o PT *vs* (4) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**
(10) w/o PT *vs* (5) w/ PT	1	**0.002**	**< 0.001**	(10) w/o PT *vs* (5) w/o PT	**< 0.001**	**< 0.001**	**< 0.001**	(10) w/o PT *vs* (6) w/ PT	0.308	1	**< 0.001**
(10) w/o PT *vs* (6) w/o PT	1	1	**0.035**								

When combined with the classic DA, each GAN-based DA remarkably outperforms the GAN-based DA or classic DA alone in terms of sensitivity since they are mutually-complementary: the former learns the non-linear manifold of the real images to generate novel local tumor features (since we train tumor/non-tumor images separately) strongly associated with sensitivity; the latter learns the geometrically-transformed manifold of the real images to cover global features and provide the robustness on training for most cases. We confirm that test samples, originally-misclassified but correctly classified after DA, are obviously different for the GAN-based DA and classic DA; here, both image-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

TABLE 7: Visual Turing Test results by an expert physician for classifying Real (R) *vs* Synthetic (S) images and Tumor (T) *vs* Non-tumor (N) images. Accuracy denotes the physician's successful classification ratio between the real/synthetic images and between the tumor/non-tumor images, respectively. It should be noted that proximity to 50% of accuracy indicates superior performance (chance = 50%).

PGGAN	Accuracy (Real *vs* Synthetic)	R as R	R as S	S as R	S as S
	79.5%	73%	27%	14%	86%
	Accuracy (Tumor *vs* Non-tumor)	T as T	T as N	N as T	N as N
	87.5%	77%	23% (R : 11, S : 12)	2% (S : 2)	98%
MUNIT	Accuracy (Real *vs* Synthetic)	R as R	R as S	S as R	S as S
	77.0%	58%	42%	4%	96%
	Accuracy (Tumor *vs* Non-tumor)	T as T	T as N	N as T	N as N
	92.5%	88%	12% (R : 6, S : 6)	3% (R : 1, S : 2)	97%
SimGAN	Accuracy (Real *vs* Synthetic)	R as R	R as S	S as R	S as S
	76.0%	53%	47%	1%	99%
	Accuracy (Tumor *vs* Non-tumor)	T as T	T as N	N as T	N as N
	94.0%	91%	9% (R : 2, S : 7)	3% (R : 3)	97%

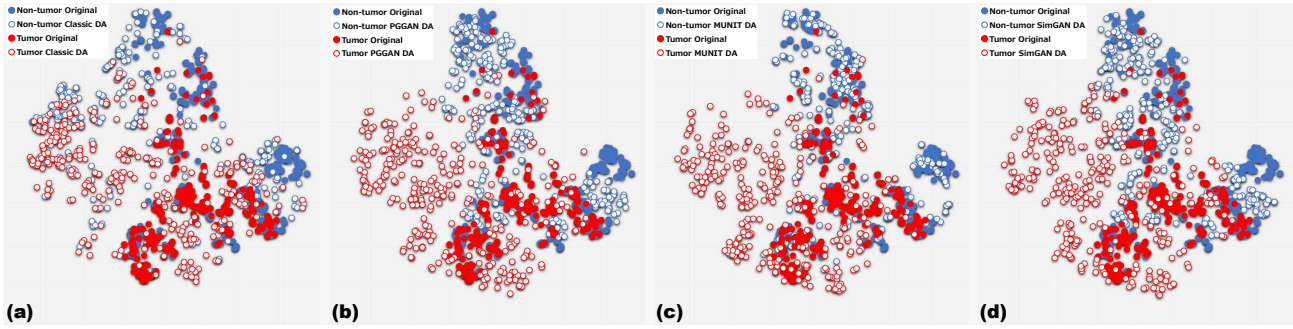

FIGURE 7: T-SNE plots with 300 tumor/non-tumor MR images per each category: Real images *vs* (a) Geometrically-transformed images; (b) PGGAN-generated images; (c) MUNIT-refined images; (d) SimGAN-refined images.

to-image GAN-based DA, especially MUNIT, produce remarkably higher sensitivity than the PGGAN-based DA after refinement. Specificity is higher than sensitivity for every DA setup with pre-training, probably due to the training data imbalance; but interestingly, without pre-training, sensitivity is higher than specificity for both image-to-image GAN-based DA since our tumor detection-oriented two-step GAN-based DA can fill the real tumor image distribution uncovered by the original dataset under no ImageNet initialization. Accordingly, when combined with the classic DA, the MUNIT-based DA based on both GANs/VAEs achieves the highest sensitivity 97.48% against the best performing classic DA's 93.67%, allowing to significantly alleviate the risk of overlooking the tumor diagnosis; in terms of sensitivity, it outperforms all the other DA setups, including two-step DA setups, with statistical significance.

D. VISUAL TURING TEST RESULTS

Table 7 indicates the confusion matrix for the Visual Turing Test. The expert physician classifies a few PGGAN-generated images as real, thanks to their realism, despite high resolution (i.e., 224×224 pixels); meanwhile, the expert classifies less GAN-refined images as real due to slight artifacts induced during refinement. The synthetic images successfully capture tumor/non-tumor features; unlike the non-tumor images, the expert recognizes a considerable number of the mild/modest tumor images as non-tumor for both real/synthetic cases. It derives from clinical tumor diagnosis relying on a full 3D volume, instead of a single 2D slice.

E. T-SNE RESULTS

As Fig. 7 represents, the real tumor/non-tumor image distributions largely overlap while the non-tumor images distribute wider. The geometrically-transformed tumor/non-tumor image distributions also often overlap, and both images distribute wider than the real ones. All GAN-based synthetic images by PGGANs/MUNIT/SimGAN distribute widely, while their tumor/non-tumor images overlap much less than the geometrically-transformed ones (i.e., a high discrimination ability associated with sensitivity improvement); the MUNIT-refined images show better tumor/non-tumor discrimination and a more similar distribution to the real ones than the PGGAN/SimGAN-based images. This trend derives from the MUNIT's loss function adopting both GANs/VAEs that further fits the PGGAN-generated images into the real image distribution by refining their texture/shape; contrarily, this refinement could also induce slight human-recognizable but DA-irrelevant artifacts. Overall, the GAN-based images, especially the MUNIT-refined images, fill the distribution uncovered by the real or geometrically-transformed ones with less tumor/non-tumor overlap; this demonstrates the superiority of combining classic DA and GAN-based DA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947606, IEEE Access

V. CONCLUSION

Visual Turing Test and t-SNE results show that PGGANs, multi-stage noise-to-image GAN, can generate realistic/diverse 256×256 brain MR images with/without tumors separately. Unlike classic DA that geometrically covers global features and provides the robustness on training for most cases, the GAN-generated images can non-linearly cover local tumor features with much less tumor/non-tumor overlap; thus, combining them can significantly boost tumor detection sensitivity—especially after refining them with MUNIT or SimGAN, image-to-image GANs; thanks to an ensemble generation process from those GANs' different algorithms, the texture/shape-refined images can replace missing data points of the training set with less tumor/non-tumor overlap, and thus handle the data imbalance by regularizing the model (i.e., improved generalization). Notably, MUNIT remarkably outperforms SimGAN in terms of sensitivity, probably due to the effect of combining both GANs/VAEs.

Regarding better medical GAN-based DA, ImageNet pre-training generally improves classification despite different textures of natural/medical images; but, without pre-training, the GAN-refined images may help achieve better sensitivity, allowing to alleviate the risk of overlooking the tumor diagnosis—this attributes to our tumor detection-oriented two-step GAN-based DA's high discrimination ability to fill the real tumor image distribution under no ImageNet initialization. GAN-generated images typically include odd artifacts; however, only without pre-training, discarding them boosts DA performance.

Overall, by minimizing the number of annotated images required for medical imaging tasks, the two-step GAN-based DA can shed light not only on classification, but also on object detection [49] and segmentation [50]. Moreover, other potential medical applications exist: (*i*) A data anonymization tool to share patients' data outside their institution for training without losing detection performance [50]; (*ii*) A physician training tool to show random pathological images for medical students/radiology trainees despite infrastructural/legal constraints [51]. As future work, we plan to define a new end-to-end GAN loss function that explicitly optimizes the classification results, instead of optimizing visual realism while maintaining diversity by combining the state-of-the-art noise-to-image and image-to-image GANs; towards this, we might extend a preliminary work on a three-player GAN for classification [52] to generate only hard-to-classify samples to improve classification; we could also (*i*) explicitly model deformation fields/intensity transformations and (*ii*) leverage unlabelled data during the generative process [53] to effectively fill the real image distribution.

ACKNOWLEDGMENT

This research was partially supported by Qdai-jump Research Program, JSPS KAKENHI Grant Number JP17K12752, and AMED Grant Number JP18lk1010028.

REFERENCES

[1] M. Havaei, A. Davy, D. Warde-Farley, et al., "Brain tumor segmentation with deep neural networks," Med. Image Anal., vol. 35, pp. 18–31, 2017.
[2] L. Rundo, C. Han, Y. Nagano, et al., "USE-Net: incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets," Neurocomputing, vol. 365, pp. 31–43, 2019.
[3] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," IEEE Access, vol. 6, pp. 9375–9389, 2017.
[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, 2015.
[5] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in Proc. International Conference on 3D Vision (3DV), pp. 565–571, 2016.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.
[7] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: a review," Med. Image Anal., 2019. (In press)
[8] C. Han, L. Rundo, R. Araki, et al., "Infinite brain MR images: PGGAN-based data augmentation for tumor detection," Neural Approaches to Dynamics of Signal Exchanges, Smart Innovation, Systems and Technologies, vol. 151, Springer, 2019. (In press)
[9] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling GANs for data augmentation in mammogram classification," Image Analysis for Moving Organ, Breast, and Thoracic Images, pp. 98–106, 2018.
[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1710.10196, 2017.
[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-Resnet and the impact of residual connections on learning," in Proc. AAAI Conference on Artificial Intelligence (AAAI), pp. 4278–4284, 2017.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
[13] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in Proc. European Conference on Computer Vision (ECCV), pp. 172–189, 2018.
[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1312.6114, 2013.
[15] A. Shrivastava, T. Pfister, O. Tuzel, et al., "Learning from simulated and unsupervised images through adversarial training," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2107–2116, 2017.
[16] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.
[17] T. Salimans, I. Goodfellow, W. Zaremba, et al., "Improved techniques for training GANs," in Advances in Neural Information Processing Systems (NIPS), pp. 2234–2242, 2016.
[18] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.
[19] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2223–2232, 2017.
[20] I. Gulrajani, F. Ahmed, M. Arjovsky, et al., "Improved training of Wasserstein GANs," in Advances in Neural Information Processing Systems (NIPS), pp. 5769–5779, 2017.
[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1511.06434, 2016.
[22] T. Xu, P. Zhang, Q. Huang, et al., "Attngan: fine-grained text to image generation with attentional generative adversarial networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1316–1324, 2018.
[23] M. Frid-Adar, I. Diamant, E. Klang, et al., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321–331, 2018.

[24] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest X-ray generation and data augmentation for cardiovascular abnormality classification," in Proc. SPIE Medical Imaging, vol. 10574, pp. 105741M, 2018.

[25] A. Gupta, S. Venkatesh, S. Chopra, and C. Ledig, "Generative image translation for data augmentation of bone lesion pathology," in Proc. International Conference on Medical Imaging with Deep Learning (MIDL), arXiv preprint arXiv:1902.02248, 2019.

[26] B. H. Menze, A. Jakab, S. Bauer, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Trans. Med. Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.

[27] S. Koley, A. K. Sadhu, P. Mitra, et al., "Delineation and diagnosis of brain tumors from post contrast T1-weighted MR images using rough granular computing and random forest," Appl. Soft Comput., vol. 41, pp. 453–465, 2016.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NIPS), pp. 1097-1105, 2012.

[29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1412.6980, 2015.

[30] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in Advances in Neural Information Processing Systems (NIPS), pp. 700–708, 2017.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1409.1556, 2015.

[32] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6924–6932, 2017.

[33] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 1501–1510, 2017.

[34] X. Mao, Q. Liu, H. Xie, R.Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2794–2802, 2017.

[35] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proc. International Conference on Computational Statistic (COMPSTAT), pp. 177–186, 2010.

[36] S. Bianco, R. Cadène, L. Celona, et al., "Benchmark analysis of representative deep neural network architectures," IEEE Access, vol. 6, pp. 64270–64277, 2018.

[37] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," Exp. Dermatol., vol. 27, no. 11, pp. 1261–1267, 2018.

[38] K. Chang, H. X. Bai, H. Zhou, et al., "Residual convolutional neural network for the determination of IDH status in low-and high-grade gliomas from MR imaging," Clin. Cancer Research, vol. 24, no. 5, pp. 1073–1081, 2018.

[39] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI," J. Magn. Reson. Imaging, vol. 49, no. 4, pp. 939–954, 2019.

[40] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

[41] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Statist., vol. 6, no. 2, pp. 65–70, 1979.

[42] R. Geirhos, P. Rubisch, C. Michaelis, et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1811.12231, 2019.

[43] F. Iandola, M. Moskewicz, S. Karaayev, et al., "DenseNet: implementing efficient ConvNet descriptor pyramids," arXiv preprint arXiv:1404.1869, 2014.

[44] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in Proc. Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027–1035, 2007.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, et al., "Dropout: a simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[46] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in Proc. International Conference on Machine Learning (ICML), vol. 37, pp. 448–456, 2015.

[47] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 240–244, 2018.

[48] C. Han, H. Hayashi, L. Rundo, et al., "GAN-based synthetic brain MR image generation," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 734–738, 2018.

[49] C. Han, K. Murao, T. Noguchi, et al., "Learning more with less: conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images," arXiv preprint arXiv:1902.09856, 2019.

[50] H. C. Shin, N. A. Renenholtz, J. K. Rogers, et al., "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in Proc. International Workshop on Simulation and Synthesis in Medical Imaging, pp. 1–11, 2018.

[51] S. G. Finlayson, H. Lee, I. S. Kohane, and L. Oakden-Rayner, "Towards generative adversarial networks as a new paradigm for radiology education," in Proc. Machine Learning for Health (ML4H) Workshop, arXiv preprint arXiv:1812.01547, 2018.

[52] S. Vandenhende, B. De Brabandere, D. Neven, and L. Van Gool, "A three-player GAN: generating hard samples to improve classification networks," in Proc. International Conference on Machine Vision Applications (MVA), arXiv preprint arXiv:1903.03496, 2019.

[53] K. Chaitanya, N. Karani, C. F. Baumgartner, et al., "Semi-supervised and task-driven data augmentation," in Proc. International Conference on Information Processing in Medical Imaging (IPMI), pp. 29–41, 2019.