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

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

[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 Centre, 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]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) can achieve excellent computer-assisted diagnosis performance, relying on sufficient annotated training data. Unfortunately, most medical imaging datasets, often collected from various scanners, are small and fragmented. In this context, as a Data Augmentation (DA) technique, 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 images 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 (*i*) noise-to-image GANs and image-to-image GANs or (*ii*) GANs and other deep generative models, 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 MR images with/without tumors separately: (*i*) Progressive Growing of GANs (PGGANs), multi-stage noise-to-image GAN for high-resolution image generation, first generates realistic/diverse 256×256 images—even a physician cannot accurately distinguish them from real ones *via* Visual Turing Test; (*ii*) UNsupervised Image-to-image Translation or SimGAN, image-to-image GAN combining GANs/Variational AutoEncoders or using a GAN loss for DA, 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 from 93.63% to 97.53%) and also in other medical imaging tasks.

INDEX TERMS Data augmentation, Synthetic image generation, GAN, 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.

The main problem in computer-assisted diagnosis lies in small and fragmented medical imaging datasets from various scanners; thus, researchers have improved classification by augmenting images with noise-to-image GANs (e.g., random noise samples to diverse pathological images [7]) or image-to-image GANs (e.g., a benign image to a malignant one [8]). However, no research has reported results achieved by combining (*i*) noise-to-image GANs and image-to-image GANs or (*ii*) GANs and other common deep generative models, such as Variational AutoEncoders (VAEs) using a single objective [9], for further performance boost.

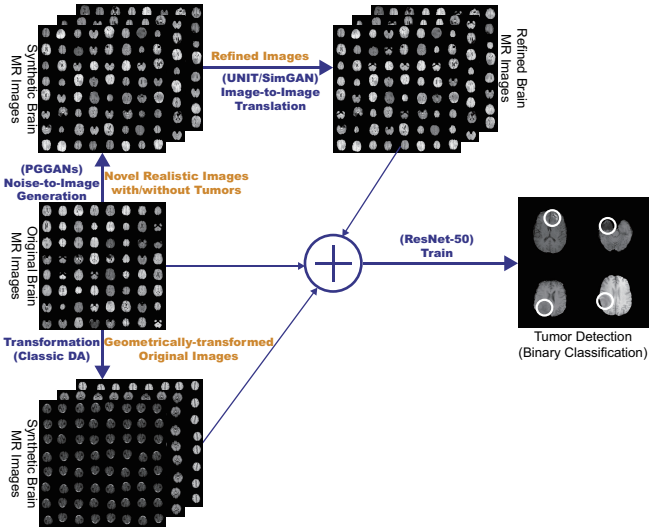

FIGURE 1: Combining noise-to-image and image-to-image GAN-based DA for better tumor detection: the PGGANs generates a number of realistic brain tumor/non-tumor MR images separately, the UNIT/SimGAN refines them separately, and the binary classifier uses them as additional training data.

So, how can we maximize DA effect under limited training images using the GAN combinations? Aiming to generate and refine brain 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 and diverse 256×256 images—the PGGANs is beneficial for DA since most CNN architectures adopt around 256×256 input sizes (e.g., InceptionResNetV2 [11]: 299×299, ResNet-50 [12]: 224×224); (*ii*) UNsupervised Image-to-image Translation (UNIT) [13] or SimGAN [14], image-to-image GAN combining GANs/VAEs or using a GAN loss for DA, further refines the texture/shape of the PGGAN-generated images to fit them into the real image distribution. We thoroughly investigate CNN-based tumor classification results, also considering the influence of pre-training on ImageNet [15] and discarding weird-looking GAN-generated images. Moreover, we evaluate the synthetic images' realism *via* Visual Turing Test [16] by an expert physician, and visualize the data distribution of real/synthetic images *via* t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [17]. When combined with classic DA, our two-step GAN-based DA approach remarkably outperforms the classic DA alone, boosting sensitivity 93.63% to 97.53%[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 [7] that aimed at investigating the potential of the PGGANs pre-trained on ImageNet—with minimal pre-processing and no refinement—for DA using a vanilla version of ResNet-50 (i.e., neither hyper-parameters nor settings were optimized).

Contributions. Our main contributions are as follows:

- **Whole Image Generation:** This research shows that PGGANs can generate realistic/diverse 256×256 whole medical images, and not only small pathological areas.
- **Two-step GAN-based DA:** This novel two-step approach, combining for the first time noise-to-image and image-to-image GANs, remarkably boosts tumor detection performance.
- **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 from various scanners.

II. GENERATIVE ADVERSARIAL NETWORKS

VAEs often suffer from blurred samples despite easier training, due to the injected noise and imperfect reconstruction using a single objective function; meanwhile, GANs [6] have revolutionized image generation in terms of realism and diversity [18] 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 and the generator's synthetic images. However, difficult GAN training from the two-player objective function accompanies artifacts and mode collapse [19], when generating high-resolution images (e.g., 256×256 pixels) [20]; to tackle this, multi-stage noise-to-image GANs have been proposed: AttnGAN [21] generates images from text using attention-based multi-stage refinement; PGGANs [10] generates realistic images using incremental multi-stage training from low resolution to high. Contrarily, to obtain images with desired texture and shape, researchers have proposed image-to-image GANs: UNIT [13] translates images using both GANs and VAEs; SimGAN [14] translates images for DA using a self-regularization term and 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) [22] and chest cardiovascular abnormality X-ray images [23]; others used the image-to-image GANs to augment breast cancer mammography images [8] and bone lesion X-ray images [24], translating benign images to 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 areas) alone, for robust classification. Along with classic image transformations, a novel approach—augmenting realistic and diverse whole medical images with the two-step GAN—may become a clinical breakthrough.

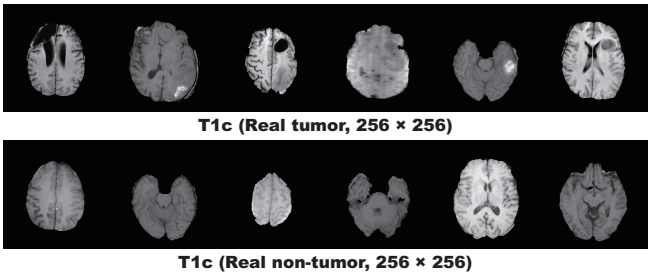

FIGURE 2: Example real MR images used for PGGAN training.

III. MATERIALS AND METHODS

A. BRATS 2016 TRAINING DATASET

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 [25]. T1c is the most common sequence in tumor detection thanks to its high-contrast [26].

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 GAN 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.

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 and 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 using gradient penalty [19]:

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

where the discriminator D is the set of 1-Lipschitz functions, $\mathbb{P}_r$ is the data distribution by the true data sample y, and $\mathbb{P}_g$ is the model distribution by the generated sample $\tilde{y}$. A gradient penalty is added for the random sample $\hat{y} \sim \mathbb{P}_{\hat{y}}$.

We train it for 100 epochs with a batch size of 16 and 1.0×10^{-3} learning rate for the Adam optimizer [27]. During training, we apply random cropping in 0–15 pixels as DA.

[2]Although this discarding procedure could be automated, we manually conducted it for more reliability; this does not affect our conclusion.

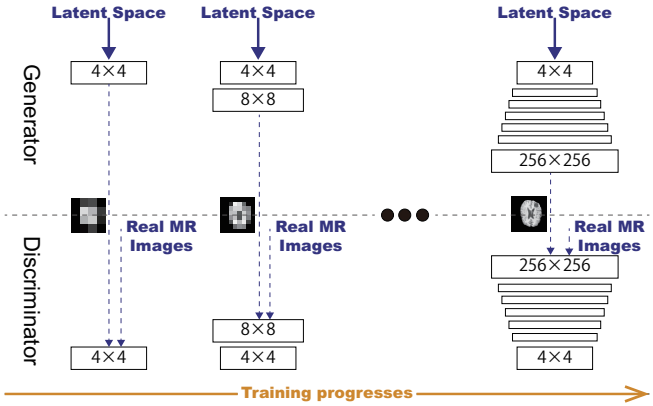

FIGURE 3: PGGAN architecture for 256×256 image generation.

C. UNIT/SIMGAN-BASED IMAGE REFINEMENT

Refinement We further refine the texture and shape of PGGAN-generated tumor/non-tumor images separately to fit them into the real image distribution using UNIT [13] or SimGAN [14]. SimGAN remarkably improved eye gaze estimation results after refining non-GAN-based synthetic images from the UnityEyes simulator *via* image-to-image translation [14]; 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., UNIT/SimGAN) with considerably different GAN-based algorithms.

We randomly select 3,000 real/3,000 PGGAN-generated tumor images for tumor image training, and we performed the same for non-tumor image training. To find suitable refining steps for each architecture, we pick the UNIT/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.

UNIT [13] is an image-to-image translation method based on both GANs and VAEs; it jointly learns image distributions in different domains using images from the marginal distributions in each domain with a shared-latent space.

UNIT Implementation Details The UNIT architecture adopts the following loss:

$$\begin{aligned}\min_{E_1,E_2,G_1,G_2}\max_{D_1,D_2} \quad & \mathcal{L}_{\text{VAE}_1}(E_1,G_1) + \mathcal{L}_{\text{GAN}_1}(E_2,G_1,D_1) \\ + \quad & \mathcal{L}_{\text{CC}_1}(E_1,G_1,E_2,G_2) \\ & \mathcal{L}_{\text{VAE}_2}(E_2,G_2) + \mathcal{L}_{\text{GAN}_2}(E_1,G_2,D_2) \\ + \quad & \mathcal{L}_{\text{CC}_2}(E_2,G_2,E_1,G_1). \end{aligned} \quad (2)$$

Using the multiple encoders E_1/E_2, generators G_1/G_2, discriminators D_1/D_2, and cycle-consistencies CC_1/CC_2, it

jointly solves learning problems of the VAE_1/VAE_2 and GAN_1/GAN_2 for the image reconstruction streams, image translation streams, and cycle-reconstruction streams.

We train it for $100,000$ steps with a batch size of 1 and 1.0×10^{-4} learning rate for the Adam optimizer [27]. The learning rate is reduced by half every 20,000 steps. During training, we apply horizontal flipping as DA.

SimGAN [14] is an image-to-image GAN designed for DA that adopts a self-regularization term/local adversarial loss; it updates a discriminator with a history of refined images.

SimGAN Implementation Details The SimGAN architecture adopts the following loss:

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

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

We train it for $20,000$ steps with a batch size of 10 and 1.0×10^{-4} learning rate for the Stochastic Gradient Descent (SGD) optimizer [28]. The learning rate is reduced by half at 15,000 steps. During training, we apply horizontal flipping as DA. We use batch normalization [29] layers.

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 synthetic images from 256×256.

ResNet-50 [12] is a 50-layer residual learning-based CNN and we adopt it to detect brain tumors in MR images (i.e., the binary classification of images with/without tumors). We chose the ResNet-50 for comparing DA setups due to its outstanding performance in image classification tasks [30].

To confirm the effect of PGGAN-based DA and its refinement using UNIT/SimGAN, we compare the following 10 DA setups under sufficient images both with/without ImageNet [15] pre-training/fine-tuning (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 UNIT-refined DA;
8) + 200k classic DA & 200k UNIT-refined DA;
9) + 200k SimGAN-refined DA;
10) + 200k classic DA & 200k SimGAN-refined DA.

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 textures rather than shapes [31]; thus, we aim to investigate how the

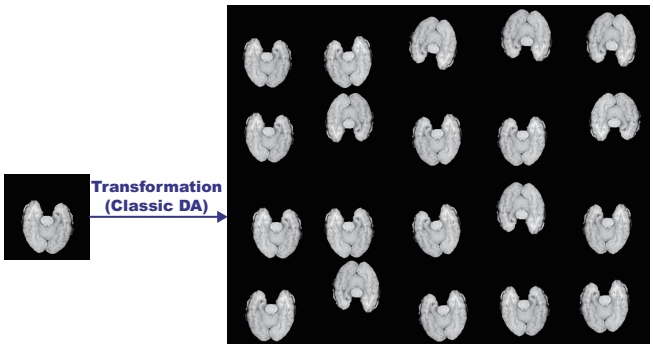

FIGURE 4: Example real MR image and its geometrically-transformed images.

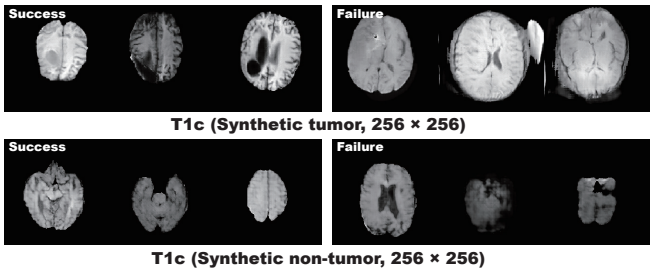

FIGURE 5: Example PGGAN-generated MR images: (a) success cases; (b) failed cases.

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 [32] extracts image features and k-means++ [33] 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 the 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. For robust training, before the final sigmoid layer, we use a 0.5 dropout [34], linear dense, and batch normalization [29] 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 [28] 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.

E. CLINICAL VALIDATION USING VISUAL TURING TEST

To quantitatively evaluate the (*i*) realism of the PGGAN-based synthetic images and (*ii*) clearness of their tumor/non-tumor features, we supply, in random order, to an expert

TABLE 1: ResNet-50 tumor detection (i.e., binary classification) results with various DA, with (without) ImageNet pre-training.

	Accuracy (%)	**Sensitivity** (%)	**Specificity** (%)
8,429 real images	93.26 (86.38)	90.95 (88.94)	95.87 (83.62)
+ 200k classic DA	95.02 (92.21)	93.63 (90.21)	96.57 (95.11)
+ 400k classic DA	94.93 (93.24)	91.90 (90.91)	98.39 (95.97)
+ 200k PGGAN-based DA	93.95 (86.25)	92.48 (87.25)	95.56 (84.78)
+ 200k PGGAN-based DA w/o clustering/discarding	94.80 (80.54)	91.82 (80.02)	98.39 (81.25)
+ 200k classic DA & 200k PGGAN-based DA	96.18 (95.63)	94.12 (94.24)	**98.79** (97.28)
+ 200k UNIT-refined DA	94.31 (83.68)	93.26 (87.75)	96.02 (78.48)
+ 200k classic DA & 200k UNIT-refined DA	**96.70** (96.34)	95.48 (**97.53**)	98.29 (94.96)
+ 200k SimGAN-refined DA	94.49 (77.66)	92.39 (82.03)	97.18 (71.98)
+ 200k classic DA & 200k SimGAN-refined DA	96.36 (95.04)	95.11 (95.07)	97.88 (94.96)

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 [16] can probe the human ability to identify attributes and relationships in images, also for visually evaluating GAN-generated images [14]; this also applies to medical images for clinical decision-making tasks [35], [36], wherein physicians' expertise is critical.

F. VISUALIZATION USING T-SNE

To visually analyze distributions of geometrically-transformed and each GAN-based images by PGGANs/UNIT/SimGAN against real images (i.e., 4 setups), we adopt t-SNE [17] 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.

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.

IV. RESULTS

This section shows how PGGANs generates synthetic brain MR images and how UNIT 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.

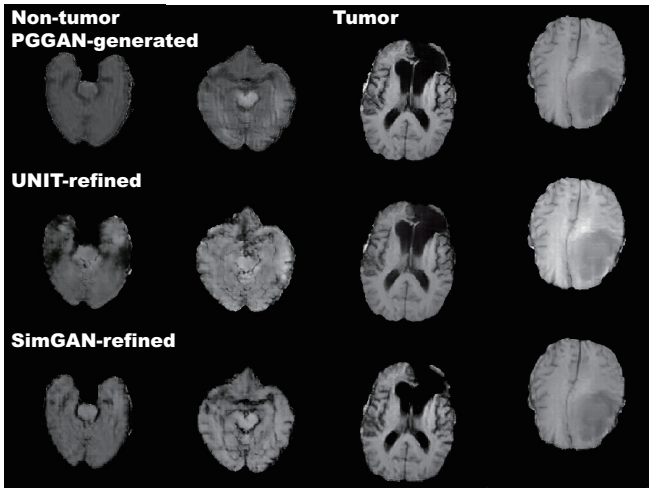

FIGURE 6: Example PGGAN-generated MR images and their refined versions by UNIT/SimGAN.

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 UNIT/SIMGAN

UNIT and SimGAN differently refine PGGAN-generated images—they render the texture/contours while maintaining the overall shape (Fig. 6). Non-tumor images change more remarkably than tumor images for both UNIT/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 1 shows the brain tumor classification results with/without DA. ImageNet pre-training generally outper-

TABLE 2: Visual Turing Test results by an expert physican for classifying Real (R) *vs* PGGAN-based Synthetic (S) images and Tumor (T) *vs* Non-tumor (N) images.

Real/Synthetic Classification	R as R	R as S	S as R	R as R
79.5%	73	27	14	86
Tumor/Non-tumor Classification	**T as T**	**T as N**	**N as T**	**N as N**
87.5%	77	23 (*R*: 11, *S*: 12)	2 (*S*: 2)	98

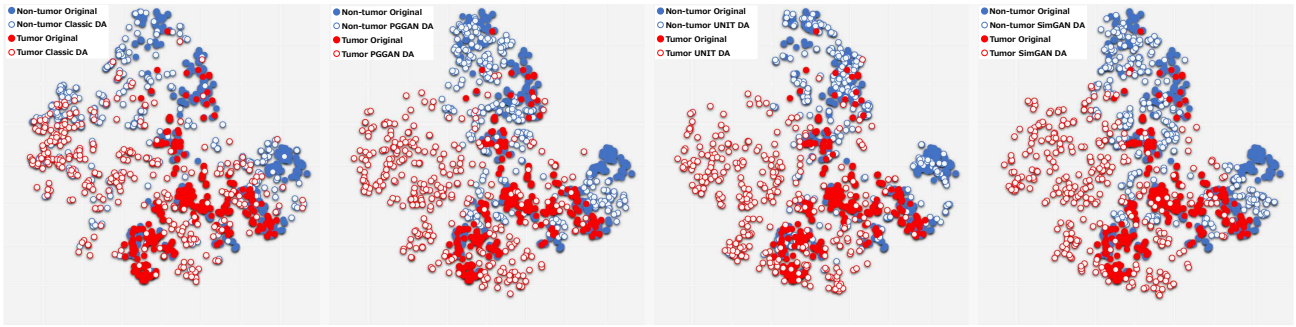

FIGURE 7: T-SNE plots with 300 tumor/non-tumor MR images per each category: Real images *vs* (a) Geometrically-transformed images; (b) PGGAN-generated images; (c) UNIT-refined images; (d) SimGAN-refined images.

forms 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/UNIT/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.

When combined with the classic DA, each GAN-based DA significantly outperforms the GAN-based DA or classic DA alone—the former fills the real image distribution uncovered by the original dataset, while the latter provides the robustness on training for most cases; here, both image-to-image GAN-based DA, especially UNIT, 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—thus, when combined with the classic DA, the UNIT-based DA achieves the highest sensitivity 97.53%, allowing to significantly alleviate the risk of overlooking the tumor diagnosis.

D. VISUAL TURING TEST RESULTS

Table 2 indicates the confusion matrix for the Visual Turing Test. The expert physician classifies a few PGGAN-generated images as real despite their high resolution (i.e., 256×256 pixels). 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/UNIT/SimGAN distribute widely, while their tumor/non-tumor images overlap much less than the geometrically-transformed ones; the UNIT-refined images show a more similar distribution to the real ones than the PGGAN/SimGAN-based images, probably due to the UNIT's loss function adopting both GANs/VAEs—overall, the GAN-based images, especially the UNIT-refined images, fill the distribution uncovered by the real or geometrically-transformed ones with less tumor/non-tumor overlap.

V. CONCLUSION

Visual Turing Test and t-SNE results show that PGGANs, multi-stage noise-to-image GAN, can generate realistic and diverse 256×256 brain MR images with/without tumors separately. The generated images can improve tumor classification, when combined with classic DA—especially af-

ter refining them with UNIT or SimGAN, image-to-image GANs; thanks to an ensemble effect from those GANs' different algorithms, the refined images can replace missing data points of the training dataset with less tumor/non-tumor overlap and regularize the model, and thus handle the data imbalance with improved generalization. Especially, UNIT outperforms SimGAN, probably due to the effect of combining both GANs and 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. 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 [37] and segmentation [38]. Moreover, other potential medical applications exist: (*i*) A data anonymization tool to share patients' data outside their institution for training without losing detection performance. This GAN-based application is reported in [38]; (*ii*) A physician training tool to show random pathological images for medical students/radiology trainees despite infrastructural/legal constraints [39]. As future work, we plan to define a new GAN loss function that explictly aims at optimizing the classification results, instead of visual realism, similarly to the three-player GAN proposed in [40].

VI. 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," arXiv preprint arXiv:1904.08254, 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] 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, Springer, arXiv preprint arXiv:1903.12564, 2019 (In press).

[8] 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.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1312.6114, 2013.

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.
[32] F. Iandola, M. Moskewicz, S. Karaayev, et al., "DenseNet: Implementing efficient ConvNet descriptor pyramids," arXiv preprint arXiv:1404.1869, 2014.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] 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.

CHANGHEE HAN received the Bachelor's and Master's Degrees in Computer Science from The University of Tokyo in 2015 and 2017, respectively. Since 2017, he is a Ph.D. student at Graduate School of Information Science and Technology, The University of Tokyo. He is currently a Visiting Scholar at National Center for Global Health and Medicine since 2018. He was invited as a Visiting Scholar at the Technical University of Munich in 2016 and University of Milano-Bicocca in 2017 and 2018. His research interests include Machine Learning, especially Deep Learning for Medical Imaging and Bioinformatics.

LEONARDO RUNDO received the Bachelor's and Master's Degrees in Computer Science Engineering from the University of Palermo in 2010 and 2013, respectively. In 2013, he was a Research Fellow at the Institute of Molecular Bioimaging and Physiology, National Research Council of Italy (IBFM-CNR). He obtained his Ph.D. in Computer Science at the University of Milano-Bicocca in 2019. Since November 2018, he is a Research Associate at the Department of Radiology, University of Cambridge, collaborating with Cancer Research UK. His main scientific interests include Biomedical Image Analysis, Machine Learning, Computational Intelligence, and High-Performance Computing.

RYOSUKE ARAKI received the Bachelor's and Master's Degrees in Engineering from Chubu University, in 2017 and 2019, respectively. Since 2019, he is a Ph.D. student at Graduate School of Engineering, Chubu University. His research interests include Computer Vision and Robot Vision, especially Deep Learning for Intelligent Robotics.

YUDAI NAGANO received the Master's Degree in Computer Science from the University of Tokyo in 2019. Since 2019, he is a Ph.D. student at Graduate school of Information Science and Technology, the University of Tokyo. His main research interests include Generative Adversarial Networks for super-resolution, image-to-image translation, and segmentation.

YUJIRO FURUKAWA received the M.D. from Akita University in 2015. Since 2019, he is a Psychiatrist at the Jikei University Hospital. His research interests include Medical Imaging of Dementia and Depression.

GIANCARLO MAURI is a Full Professor of Computer Science at University of Milano-Bicocca. His research interests include: natural computing and unconventional computing models, bioinformatics, stochastic modeling, and simulation of biological systems and processes using High-Performance Computing approaches, biomedical data analysis. On these subjects, he published about 400 scientific papers. He is a member of the steering committees of the International Conference on Membrane Computing, of the International Conference on Unconventional Computing and Natural Computing and, of the International workshop on Cellular Automata.

HIDEKI NAKAYAMA received the Master's and Ph.D. Degrees in Information Science from the University of Tokyo, Japan in 2008 and 2011, respectively. From 2012 to 2018, he was an Assistant Professor at the Graduate School of Information Science and Technology, the University of Tokyo, Japan. Since April 2018, he has been an Associate Professor at the same department. He is also an affiliated faculty of the International Research Center for Neurointelligence (IRCN), and a Visiting Researcher at the National Institute of Advanced Industrial Science and Technology (AIST). His research interests include generic image recognition, Natural Language Processing, and Deep Learning.

HIDEAKI HAYASHI received the Bachelor's, Master's, and Ph.D. degrees in Engineering from Hiroshima University in 2012, 2014, and 2016, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2015 to 2017. He is currently an assistant professor in the Department of Advanced Information Technology, Kyushu University. His research interests include biosignal analysis, neural networks, and machine learning.