Ph.D Thesis Short Version

Pathology-Aware Generative Adversarial Networks for Medical Image Augmentation

Changhee Han[1,2,a)] Hideki Nakayama[1,b)]

Abstract: Convolutional Neural Networks (CNNs) can play a key role in Medical Image Analysis under large-scale annotated datasets. However, preparing such massive dataset is demanding. In this context, Generative Adversarial Networks (GANs) can generate realistic but novel samples, and thus effectively cover the real image distribution. In terms of interpolation, the GAN-based medical image augmentation is reliable because medical modalities can display the human body's strong anatomical consistency at fixed position while clearly reflecting inter-subject variability; thus, we propose to use noise-to-image GANs (e.g., random noise samples to diverse pathological images) for (*i*) medical Data Augmentation (DA) and (*ii*) physician training. Regarding the DA, the GAN-generated images can improve Computer-Aided Diagnosis based on supervised learning. For the physician training, the GANs can display novel desired pathological images and help train medical trainees despite infrastructural/legal constraints. This thesis contains four GAN projects presenting such novel applications' clinical relevance in collaboration with physicians. Whereas the methods are more generally applicable, this thesis only explores a few oncological applications.

In the first project, after proposing the two applications, we demonstrate that GANs can generate realistic/diverse 128 × 128 whole brain Magnetic Resonance (MR) images from noise samples—despite difficult training, such noise-to-image GAN can increase image diversity for further performance boost. Even an expert fails to distinguish the synthetic images from the real ones in Visual Turing Test.

The second project tackles image augmentation for 2D classification. Most CNN architectures adopt around 256 × 256 input sizes; thus, we use the noise-to-noise GAN, Progressive Growing of GANs (PGGANs), to generate realistic/diverse 256 × 256 whole brain MR images with/without tumors separately. Multimodal UNsupervised Image-to-image Translation further refines the synthetic images' texture/shape. Our two-step GAN-based DA boosts sensitivity 93.7% to 97.5% in 2D tumor/non-tumor classification. An expert classifies a few synthetic images as real.

The third project augments images for 2D detection. Further DA applications require pathology localization for detection and advanced physician training needs atypical image generation, respectively. To meet both clinical demands, we propose Conditional PGGANs (CPGGANs) that incorporates highly-rough bounding box conditions incrementally into the noise-to-image GAN (i.e., the PGGANs) to place realistic/diverse brain metastases at desired positions/sizes on 256 × 256 MR images; the bounding box-based detection requires much less physicians' annotation effort than segmentation. Our CPGGAN-based DA boosts sensitivity 83% to 91% in tumor detection with acceptable False Positives (FPs). In terms of extrapolation, such pathology-aware GANs are promising because common and/or desired medical priors can play a key role in the conditioning—theoretically, infinite conditioning instances, external to the training data, exist and enforcing such constraints have an extrapolation effect *via* model reduction.

Finally, we solve image augmentation for 3D detection. Because lesions vary in 3D position/appearance, 3D multiple pathology-aware conditioning is important. Therefore, we propose 3D Multi-Conditional GAN (MCGAN) that translates noise boxes into realistic/diverse 32 × 32 × 32 lung nodules placed naturally at desired position/size/attenuation on Computed Tomography scans. Our 3D MCGAN-based DA boosts sensitivity in 3D nodule detection under any nodule size/attenuation at fixed FP rates. Considering the realism confirmed by physicians, it could perform as a physician training tool to display realistic medical images with desired abnormalities.

Two discussions confirm our pathology-aware GANs' clinical relevance: (*i*) Conducting a questionnaire survey about our GAN projects for 9 physicians; (*ii*) Holding a workshop about how to develop medical Artificial Intelligence (AI) fitting into a clinical environment in five years for 7 professionals with various AI and/or Healthcare background.

Keywords: Generative Adversarial Networks, Convolutional Neural Networks, Data Augmentation, Physician Training, Medical Image Analysis

"Life is short, and the Art long; the occasion fleeting; experience fallacious, and judgment difficult. The physician must not only be prepared to do what is right himself, but also to make the patient, the attendants, and externals cooperate."

Hippocrates [460-375 BC]

[1] The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[2] LPIXEL Inc., Chiyoda, Tokyo 100–0004, Japan
[a)] han@lpixel.net
[b)] nakayama@ci.i.u-tokyo.ac.jp

1. Introduction

Convolutional Neural Networks (CNNs) have revolutionized Medical Image Analysis, occasionally outperforming expert physicians in diagnostic accuracy when large-scale annotated datasets were available [1], [2]. However, obtaining such massive datasets often involves the following intrinsic challenges [3], [4]: (*i*) it is costly and laborious to collect medical images, such as Magnetic Resonance (MR) and Computed Tomography (CT) images, especially for rare disease; (*ii*) it is time-consuming and observer-dependent, even for expert physicians, to annotate them due to the low pathological-to-healthy ratio. To tackle these issues, researchers have mainly focused on extracting as much information as possible from the available limited data [5], [6]. Instead, Generative Adversarial Networks (GANs) [7] can generate realistic but completely new samples *via* many-to-many mappings, and thus effectively cover the real image distribution; they showed great promise in Data Augmentation (DA) [8].

Interpolation refers to new data point construction within a discretely-sampled data distribution. In terms of the interpolation, GAN-based medical image augmentation is reliable because medical modalites (e.g., X-ray, CT, MRI) can display the human body's strong anatomical consistency at fixed position while clearly reflecting inter-subject variability—this differs from the natural images, where various objects can appear at any position; accordingly, to tackle large inter-subject/pathology/modality variability [3], [4], we propose to use noise-to-image GANs (e.g., random noise samples to diverse pathological images) for (*i*) medical DA and (*ii*) physician training [9]. The noise-to-image GAN training is more challenging than training image-to-image GANs (e.g., a benign image to a malignant one); but, it can perform more global regularization (i.e., adding constraints when fitting a loss function on a training set to prevent overfitting) and increase image diversity for further performance boost.

Regarding the DA, the GAN-generated images can improve Computer-Aided Diagnosis (CAD) based on supervised learning [10]. For the physician training, the GANs can display novel desired pathological images and help train medical trainees despite infrastructural and legal constraints [11]. However, we cannot directly use conventional GANs for realistic/diverse high-resolution medical image augmentation. Moreover, we have to find effective loss functions and training schemes for each of those applications [12]; the diversity matters more for the DA to sufficiently fill the real image distribution whereas the realism matters more for the physician training not to confuse the medical students and radiology trainees.

So, how can we perform GAN-based DA/physician training using limited annotated training images? Always in collaboration with physicians, for improving 2D classification, we combine the noise-to-image [13], [14] (i.e., Progressive Growing of GANs, PGGANs [15]) and image-to-image GANs (i.e., Multimodal UNsupervised Image-to-image Translation, MUNIT [16]); the two-step GAN can generate and refine realistic/diverse 256×256 brain MR images with/without tumors separately. Nevertheless, further DA applications require pathology localization for detection (i.e., identifying target pathology positions in medical images) and advanced physician training needs atypical image generation. To meet both demands, we propose 2D/3D bounding box-based GANs conditioned on pathology position/size/appearance; the bounding box-based detection requires much less physicians' annotation effort than segmentation.

Extrapolation refers to new data point estimation beyond a discretely-sampled data distribution. While it is not mutually-exclusive with the interpolation and both rely on a model's restoring force, it is more subject to uncertainty and thus a risk of meaningless data generation. In terms of the extrapolation, the pathology-aware GANs (i.e., the conditional GANs controlling pathology, such as tumors and nodules, based on position/size/appearance) are promising because common and/or desired medical priors can play a key role in the conditioning—theoretically, infinite conditioning instances, external to the training data, exist and enforcing such constraints have an extrapolation effect *via* model reduction [17]; the reduction-caused inevitable errors, not limited between two data points, force a generator to synthesize images that it has never synthesized before.

For improving 2D detection, we propose Conditional PGGANs (CPGGANs) that incorporates highly-rough bounding box conditions incrementally into the noise-to-image GAN (i.e., the PGGANs) to place realistic/diverse brain metastases at desired positions/sizes on 256 × 256 MR images [18]. As its pathology-aware conditioning, we use 2D tumor position/size on MR images. Since lesions vary in 3D position/appearance, for improving 3D detection, we propose 3D Multi-Conditional GAN (MCGAN) that translates noise boxes into realistic/diverse 32×32×32 lung nodules placed naturally at desired position/size/attenuation on CT scans [19]; inputting the noise box with the surrounding tissues has the effect of combining the noise-to-image and image-to-image GANs. As its pathology-aware conditioning, we use 3D nodule position/size/attenuation on CT scans.

Lastly, two discussions confirm our pathology-aware GANs' clinical relevance for diagnosis as a clinical decision support system and physician training tool: (*i*) Conducting a questionnaire survey about our GAN projects for 9 physicians; (*ii*) Holding a workshop about how to develop medical Artificial Intelligence (AI) fitting into a clinical environment in five years for 7 professionals with various AI and/or Healthcare background.

Contributions. Our main contributions are as follows:

- **Noise-to-Image GAN Applications:** We propose clinically-valuable novel noise-to-image GAN applications, medical DA and physician training, focusing on their ability to generate realistic and diverse images.
- **Pathology-Aware GANs:** For required extrapolation, in collaboration with physicians, we propose novel 2D/3D GANs controlling pathology (i.e., tumors and nodules) on most major modalities (i.e., brain MRI and lung CT).
- **Clinical Validation:** After detailed discussions with many physicians and professionals with various AI and/or Healthcare background, we confirm our pathology-aware GANs' clinical relevance as a (*i*) clinical decision support system and (*ii*) non-expert physician training tool.

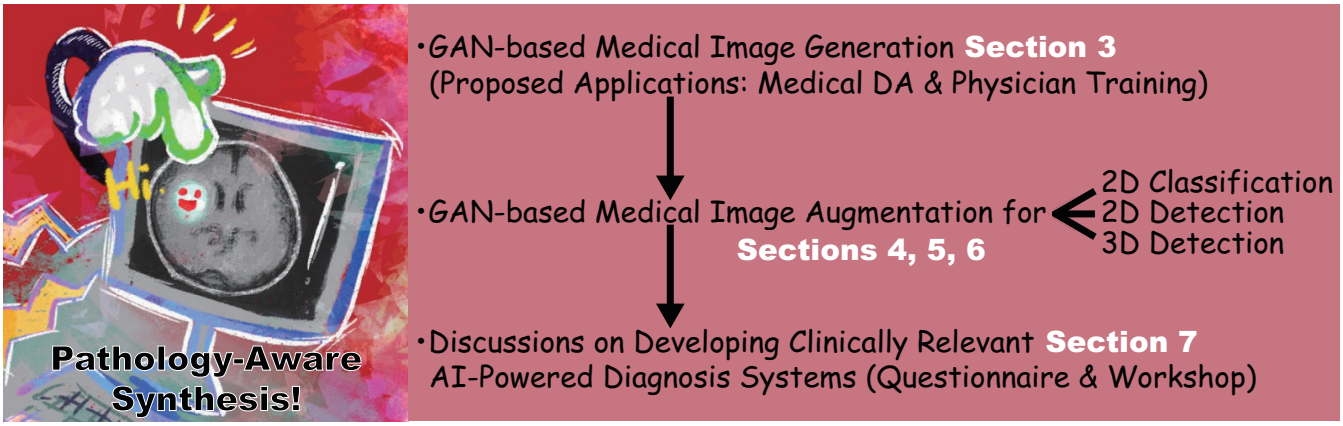

Fig. 1: Conceptual scheme of this thesis: we propose novel noise-to-image GAN-based clinical applications, (*i*) medical DA and (*ii*) physician training; then, to present such GAN applications' technical soundness, we successfully tackle 2D classification, 2D detection, and 3D detection in collaboration with physicians—we propose novel pathology-aware GANs for effective extrapolation; lastly, we discuss how to develop clinically relevant AI-powered diagnosis systems, especially focusing on our pathology-aware GAN applications, *via* a questionnaire survey and workshop.

This Ph.D. thesis aims to present the clinical relevance of our novel pathology-aware GAN applications, medical DA and physician training, always in collaboration with physicians.

The thesis is organized as follows (Fig. 1). **Section 2** describes related work on the GAN-based medical DA and physician training, which emerged after our proposal to use noise-to-image GANs for those applications in **Section 3**. **Section 4** presents a two-step GAN for 2D classification that combines both noise-to-image and image-to-image GANs. **Section 5** proposes CPG-GANs for 2D detection that incorporates highly-rough bounding box conditions incrementally into the noise-to-image GAN. Finally, we propose 3D MCGAN for 3D detection that translates noise boxes into desired pathology in **Section 6**. **Section 7** discusses both our pathology-aware GANs' clinical relevance *via* a questionnaire survey and how to develop medical AI fitting into a clinical environment in five years *via* a workshop. Lastly, **Section 8** provides the conclusive remarks and future directions for further GAN-based extrapolation.

2. Investigated Contexts and Applications

2.1 GAN-based Medical DA

Because the lack of annotated pathological images is the greatest challenge in CAD [3], [4], to handle various types of small/fragmented datasets from multiple scanners, researchers have actively conducted GAN-based DA studies especially in Medical Image Analysis. For better classification, some researchers adopted image-to-image GANs similarly to their conventional medical applications, such as denoising [20] and MRI-to-CT translation [21]: Wu *et al.* translated 256×256 normal mammograms into lesion ones [22], Gupta *et al.* translated 1024×512 normal leg X-ray images into bone lesion ones [23], and Malygina *et al.* translated $256 \times 256/512 \times 512$ normal chest X-ray images into pneumonia/pleural-thickening ones [24]. Meanwhile, others adopted the noise-to-image GANs as we proposed, to increase image diversity for further performance boost—the diversity matters more for the DA to sufficiently fill the real image distribution: Frid-Adar *et al.* augmented 64×64 liver lesion CT images [10] and Madani *et al.* augmented 128×128 chest X-ray images with cardiovascular abnormality [25].

To facilitate pathology detection and segmentation, researchers conditioned the image-to-image GANs, not the noise-to-image GANs like our work in **Section 5**, with pathology features (e.g., position, size, and appearance) and generated realistic/diverse pathology at desired positions in medical images. In terms of extrapolation, the pathology-aware GANs are promising because common and/or desired medical priors can play a key role in the conditioning—theoretically, infinite conditioning instances, external to the training data, exist and enforcing such constraints have an extrapolation effect *via* model reduction [17]. To the best of our knowledge, only Kanayama *et al.* tackled bounding box-based pathology detection using the image-to-image GAN [26]; they translated normal endoscopic images with various image sizes (458×405 on average) into gastric cancer ones by inputting both a benign image and a black image (i.e., pixel value: 0) with a specific lesion Region Of Interest (ROI) at desired position. Without conditioning the noise-to-image GAN with nodule position, Gao *et al.* generated $40 \times 40 \times 18$ 3D nodule subvolumes only applicable to their subvolume-based detector [27].

Since 3D imaging is spreading in radiology (e.g., CT, MRI), most GAN-based DA works for segmentation exploited 3D conditional image-to-image GANs. However, 3D medical image generation is more challenging than 2D one due to expensive computational cost and strong anatomical consistency; so, instead of generating a whole image including pathology, researchers only focused on a malignant Voxel Of Interest (VOI): Shin *et al.* translated $128 \times 128 \times 54$ normal brain MR images into tumor ones by inputting both a benign image and a tumor-conditioning image [28], similarly to the Kanayama *et al.*'s work [26]; Jin *et al.* generated $64 \times 64 \times 64$ CT images of lung nodules including the surrounding tissues by only inputting a VOI centered at a lung nodule, but with a central sphere region erased [29]. Recently, instead of generating realistic images and training classifiers on them separately, Chaitanya *et al.* directly optimized segmentation results on cardiac MR images [30]; however, it segmented body parts, instead of pathology. Since effective GAN-based medical DA generally requires much engineering effort, we published a tutorial journal paper [12] about tricks to improve performance using the GANs, based on our experience and related work.

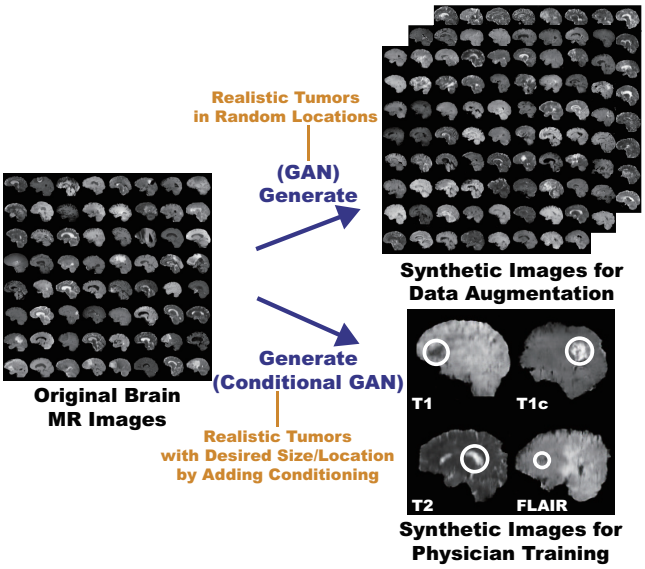

Fig. 2: Potential applications of the proposed GAN-based synthetic brain MR image generation: (1) DA for better diagnostic accuracy by generating random realistic images giving insights in classification; (2) physician training for better understanding various diseases to prevent misdiagnosis by generating desired realistic pathological images.

2.2 GAN-based Physician Training

While medical students and radiology trainees must view thousands of images to become competent [31], accessing such abundant medical images is often challenging due to infrastructural and legal constraints [32]. Because pathology-aware GANs can generate novel medical images with desired abnormalities (e.g., position, size, and appearance)—while maintaining enough realism not to confuse the medical trainees—GAN-based physician training concept is drawing attention: Chuquicusma *et al.* appreciated the GAN potential to train radiologists for educational purpose after successfully generating 56×56 CT images of lung nodules that even deceived experts [33]; thanks to their anonymization ability, Shin *et al.* proposed to share pathology-aware GAN-generated images outside institutions after achieving considerable tumor segmentation results with only synthetic $128 \times 128 \times 54$ MR images for training [28]; more importantly, Finlayson *et al.* from Harvard Medical School are currently validating a class-conditional GANs' radiology educational efficacy after succeeding in learning features that distinguish fractures from non-fractures on 1024×1024 pelvic X-ray images [11].

3. GAN-based Medical Image Generation

3.1 Motivation

How can we generate realistic medical images completely different from the original samples? Our aim is to generate synthetic multi-sequence brain MR images using GANs, which is essential in medical imaging to increase diagnostic reliability, such as *via* DA in CAD as well as physician training (Fig. 2) [34]. However, this is extremely challenging—MR images are characterized by low contrast, strong visual consistency in brain anatomy, and intra-sequence variability. Our novel GAN-based approach for medical DA adopts Deep Convolutional Generative Adversarial Network (DCGAN) [35] and WGAN [36] to generate realistic images, and an expert physician validates them *via* Visual Turing Test [37].

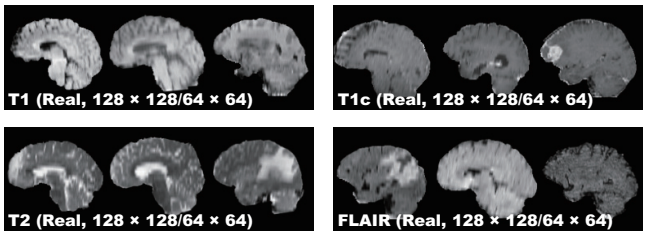

Fig. 3: Example real $128 \times 128/64 \times 64$ MR images used for GAN training: the resized sagittal multi-sequence brain MRI scans of patients with HGG on the BRATS 2016 training dataset [38].

Contributions. Our main contributions are as follows:

- **MR Image Generation:** This research shows that WGAN can generate realistic multi-sequence brain MR images, possibly leading to DA and physician training.
- **Medical Image Generation:** This research provides how to exploit medical images with intrinsic intra-sequence variability towards GAN-based DA for medical imaging.

3.2 Materials and Methods

3.2.1 BRATS 2016 Dataset

This project exploits multi-sequence 240×155 brain MR images from the the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) 2016 training dataset [38]: it contains 220 High-Grade Glioma (HGG) and 54 Low-Grade Glioma (LGG) cases, with T1-weighted (T1), contrast enhanced T1-weighted (T1c), T2-weighted, and FLAIR sequences.

3.2.2 DCGAN/WGAN-based Image Generation

Pre-processing We select the slices from #80 to #149 among the whole 240 slices to omit initial/final slices, since they convey a negligible amount of useful information and could affect the training. The images are resized to both $64 \times 64/128 \times 128$ pixels from 240×155 for better GAN training. Fig. 3 shows example real MR images used for training; each sequence contains 15,400 images with 220 patients $\times$ 70 slices (61,600 in total).

MR Image Generation DCGAN and WGAN generate six types of images as follows:

- T1 sequence (128×128) from the real T1;
- T1c sequence (128×128) from the real T1c;
- T2 sequence (128×128) from the real T2;
- FLAIR sequence (128×128) from the real FLAIR;
- Concat sequence (128×128) from concatenating the real T1, T1c, T2, and FLAIR (i.e., feeding the model with samples from all the MRI sequences);
- Concat sequence (64×64) from concatenating the real T1, T1c, T2, and FLAIR.

Concat sequence refers to a new ensemble sequence for an alternative DA, containing features of all four sequences.

DCGAN [35] is a standard GAN [7] with a convolutional architecture for unsupervised learning.

DCGAN Implementation Details We use the same DCGAN architecture [35] with no tanh in the generator, ELU as the discriminator, all filters of size 4×4, and a half channel size for DCGAN training. A batch size of 64 and Adam optimizer with 2.0×10^{-4} learning rate were implemented.

WGAN [36] is an alternative to traditional GAN training, as the JS divergence is limited, such as when it is discontinuous.

Table 1: Visual Turing Test results by a physician for classifying 50 real *vs* 50 synthetic 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%).

	Accuracy (%)	Real as Real (%)	Real as Synt (%)	Synt as Real (%)	Synt as Synt (%)
T1 (DCGAN, 128 × 128)	70	52	48	12	88
T1c (DCGAN, 128 × 128)	71	48	52	6	94
T2 (DCGAN, 128 × 128)	64	44	56	16	84
FLAIR (DCGAN, 128 × 128)	54	24	76	16	84
Concat (DCGAN, 128 × 128)	77	68	32	14	86
Concat (DCGAN, 64 × 64)	54	26	74	18	82
T1 (WGAN, 128 × 128)	64	40	60	12	88
T1c (WGAN, 128 × 128)	55	26	74	16	84
T2 (WGAN, 128 × 128)	58	38	62	22	78
FLAIR (WGAN, 128 × 128)	62	32	68	8	92
Concat (WGAN, 128 × 128)	66	62	38	30	70
Concat (WGAN, 64 × 64)	53	36	64	30	70

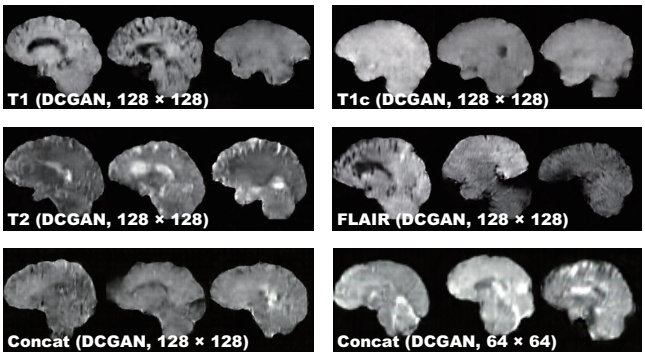

Fig. 4: Example 128 × 128/64 × 64 DCGAN-generated MR images.

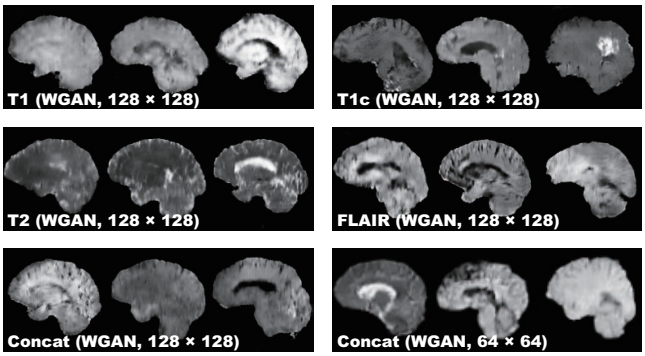

Fig. 5: Example 128 × 128/64 × 64 WGAN-generated MR images.

WGAN Implementation Details We use the same DCGAN architecture [35] for WGAN training. A batch size of 64 and Root Mean Square Propagation (RMSprop) optimizer with 5.0×10^{-5} learning rate were implemented.

3.2.3 Clinical Validation *via* Visual Turing Test

To quantitatively evaluate how realistic the synthetic images are, an expert physician was asked to constantly classify a random selection of 50 real/50 synthetic MR images as real or synthetic shown in random order for each GAN/sequence, without previous training stages revealing which is real/synthetic.

3.3 Results

3.3.1 MR Images Generated by DCGAN/WGAN

DCGAN Fig. 4 illustrates examples of synthetic images by DCGAN. The images look similar to the real samples. Concat images combine appearances and patterns from all the four sequences used in training. Since DCGAN's value function could be unstable, it often generates hyper-intense T1-like images.

WGAN Fig. 5 shows the example output of WGAN in each sequence. Remarkably outperforming DCGAN, WGAN successfully captures the sequence-specific texture and tumor appearance while maintaining the realism of the original brain MR images.

3.3.2 Visual Turing Test Results

Table 1 shows the confusion matrix concerning the Visual Turing Test. Even the expert physician found classifying real and synthetic images challenging, especially in lower resolution due to their less detailed appearances. WGAN succeeded to deceive the physician significantly better than DCGAN for all the MRI sequences except FLAIR images (62% to 54%).

3.4 Conclusion

Our results show that GANs, especially WGAN, can generate 128 × 128 realistic multi-sequence brain MR images that even a physician is unable to accurately distinguish from the real, leading to DA/physician training. This attributes to WGAN's good generalization ability with a sharp value function.

4. GAN-based Medical Image Augmentation for 2D Classification

How can we maximize the DA effect under limited training images using the GAN combinations? To generate and refine brain MR images with/without tumors separately (Fig. 6), we propose a two-step GAN-based DA approach: (*i*) PGGANs [15], low-to-high resolution noise-to-image GAN, first generates realistic/diverse 256 × 256 images—the PGGANs helps DA since most CNN architectures adopt around 256 × 256 input sizes; (*ii*) MUNIT [16] that combines GANs/Variational AutoEncoders (VAEs) [39] or SimGAN [8] that uses a DA-focused GAN loss, further refines the texture and shape of the PGGAN-generated images to fit them into the real image distribution.

4.1 Motivation

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 sub-areas—and MUNIT can further refine their texture and shape similarly to real ones.
- **Two-step GAN-based DA:** This two-step approach, combining for the first time noise-to-image and image-to-image GANs, significantly boosts tumor classification 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 and fragmented datasets.

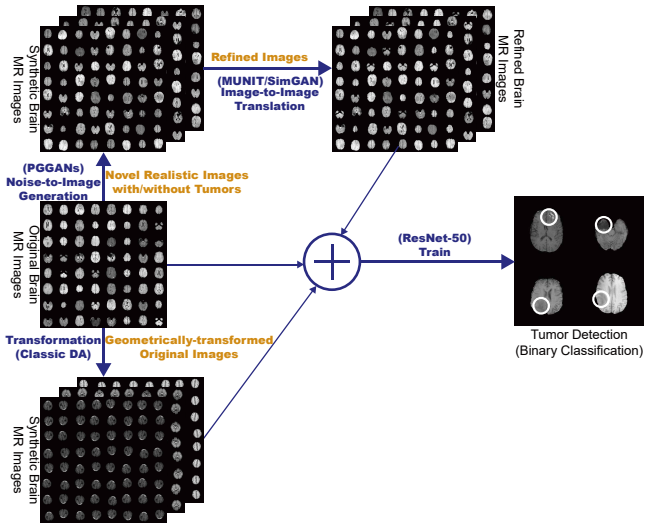

Fig. 6: Combining noise-to-image and image-to-image GANs for better tumor classification: the PGGANs generates a number of realistic brain tumor/non-tumor MR images separately, MUNIT/SimGAN refines them separately, and binary classifier uses them for training.

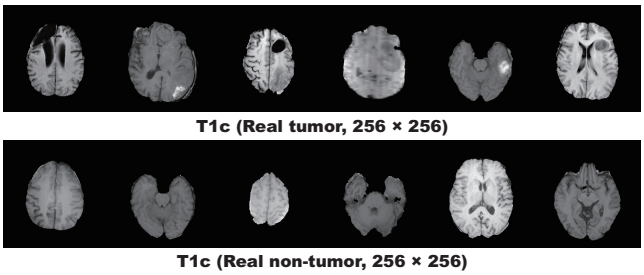

Fig. 7: Example real 256 × 256 MR images used for PGGAN training.

4.2 Materials and Methods

4.2.1 BRATS 2016 Dataset

We use a dataset of 240 × 240 T1c brain axial MR images of 220 HGG cases from BRATS 2016 [38]. T1c is the most common sequence in tumor classification thanks to its high-contrast [40].

4.2.2 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. For tumor classification, 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. 7 shows example real MR images.

PGGANs [15] is a GAN training method that progressively grows a generator and discriminator. We train and generate 256×256 tumor/non-tumor images separately with the PGGANs.

PGGAN Implementation Details The PGGAN architecture adopts the Wasserstein loss with Gradient Penalty (WGAN-GP) [41]. We train the model for 100 epochs with a batch size of 16 and 1.0×10^{-3} learning rate for the Adam optimizer [42]. During training, we apply random cropping in 0-15 pixels as DA.

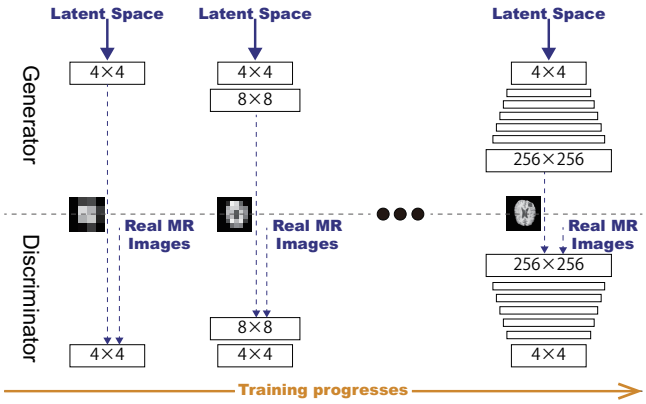

Fig. 8: PGGAN architecture for 256 × 256 brain MR image generation. $N \times N$ refers to convolutional layers operating on $N \times N$ spatial resolution.

4.2.3 MUNIT/SimGAN-based Image Refinement

Refinement Using resized 224 × 224 images for ResNet-50, we further refine the texture and shape of PGGAN-generated tumor/non-tumor images separately to fit them into the real image distribution using MUNIT [16] or SimGAN [8].

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

MUNIT [16] is an image-to-image GAN based on both auto-encoding/translation.

MUNIT Implementation Details We train the model for 100, 000 steps with a batch size of 1 and 1.0×10^{-4} learning rate (it halves every 20, 000 steps) for the Adam optimizer [42]. During training, we apply horizontal flipping as DA.

SimGAN [8] 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 We train the model for 20, 000 steps with a batch size of 10 and 1.0×10^{-4} learning rate for the Stochastic Gradient Descent (SGD) optimizer [43]. The learning rate is reduced by half at 15,000 steps. During training, we apply horizontal flipping as DA.

4.2.4 ResNet-50-based Tumor Classification

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 [44] is a 50-layer residual learning-based CNN.

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 [45] pre-training (i.e., 20 DA setups):

(1) 8, 429 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;

Table 2: ResNet-50 tumor results of 20 DA setups, with (without) ImageNet pre-training.

DA Setups	Accuracy (%)	Sensitivity (%)	Specificity (%)
(1) 8,429 real images	93.1 (86.3)	90.9 (88.9)	95.9 (83.2)
(2) + 200k classic DA	95.0 (92.2)	93.7 (89.9)	96.6 (95.0)
(3) + 400k classic DA	94.8 (93.2)	91.9 (90.9)	98.4 (96.1)
(4) + 200k PGGAN-based DA	93.9 (86.2)	92.6 (87.3)	95.6 (84.9)
(5) + 200k PGGAN-based DA w/o clustering/discarding	94.8 (80.7)	91.9 (80.2)	98.4 (81.2)
(6) + 200k classic DA & 200k PGGAN-based DA	96.2 (95.6)	94.0 (94.2)	**98.8** (97.3)
(7) + 200k MUNIT-refined DA	94.3 (83.7)	93.0 (87.8)	95.8 (78.5)
(8) + 200k classic DA & 200k MUNIT-refined DA	**96.7** (96.3)	95.4 (**97.5**)	98.2 (95.0)
(9) + 200k SimGAN-refined DA	94.5 (77.6)	92.3 (82.3)	97.1 (72.0)
(10) + 200k classic DA & 200k SimGAN-refined DA	96.4 (95.0)	95.1 (95.1)	97.9 (95.0)

(8) + 200k classic DA & 200k MUNIT-refined DA;
(9) + 200k SimGAN-refined DA;
(10)+ 200k classic DA & 200k SimGAN-refined DA.

We aim to achieve higher sensitivity, using the additional synthetic training images. This paper 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.

ResNet-50 Implementation Details The ResNet-50 architecture adopts the binary cross-entropy loss. For robust training, before the final sigmoid layer, we introduce a 0.5 dropout [46], linear dense, and batch normalization [47] layers. We use a batch size of 96, 1.0×10^{-2} learning rate for the SGD optimizer [43], 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.

4.2.5 Clinical Validation *via* Visual Turing Test

To quantify the (*i*) realism of 224×224 synthetic images by PGGANs, MUNIT, and SimGAN against real ones respectively (i.e., 3 setups) and (*ii*) clearness of their tumor/non-tumor features, we supply, in random order, to aphysician 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 is asked to classify them as both (*i*) real/synthetic and (*ii*) tumor/non-tumor.

4.2.6 Visualization *via* 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-Distributed Stochastic Neighbor Embedding (t-SNE) [48] 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 visualization.

T-SNE Implementation Details The t-SNE uses a perplexity of 100 for $1,000$ iterations to visually represent a 2D space.

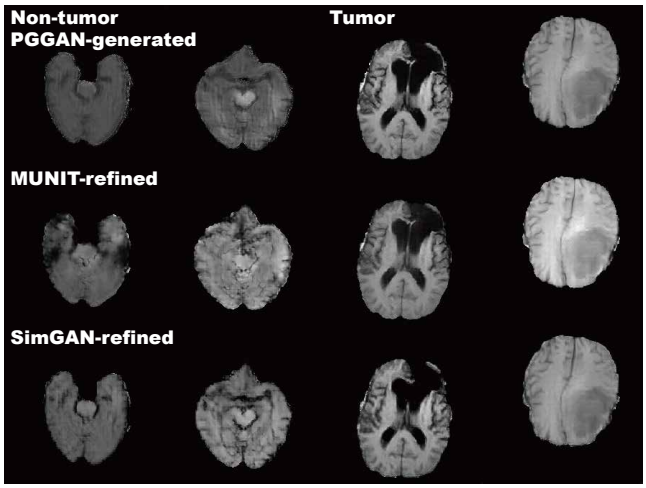

Fig. 9: Example PGGAN-generated 256×256 MR images and their refined versions by MUNIT/SimGAN.

4.3 Results

4.3.1 MR Images Generated by PGGANs

Fig. 9 illustrates examples of synthetic MR images by PGGANs. For about 75% of cases, it successfully captures the T1c-specific texture and tumor appearance, while maintaining the realism; but, for the rest 25%, the generated images lack clear tumor/non-tumor features or contain unrealistic features.

4.3.2 MR Images Refined by MUNIT/SimGAN

MUNIT and SimGAN differently refine PGGAN-generated images' texture/shape (Fig. 9).

4.3.3 Tumor Classification Results

Table 2 shows the tumor classification results with/without DA. 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.

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

Table 3: Visual Turing Test results by an expert physician for classifying Real (R) *vs* Synthetic (S) images/ Tumor (T) *vs* Non-tumor (N) images.

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

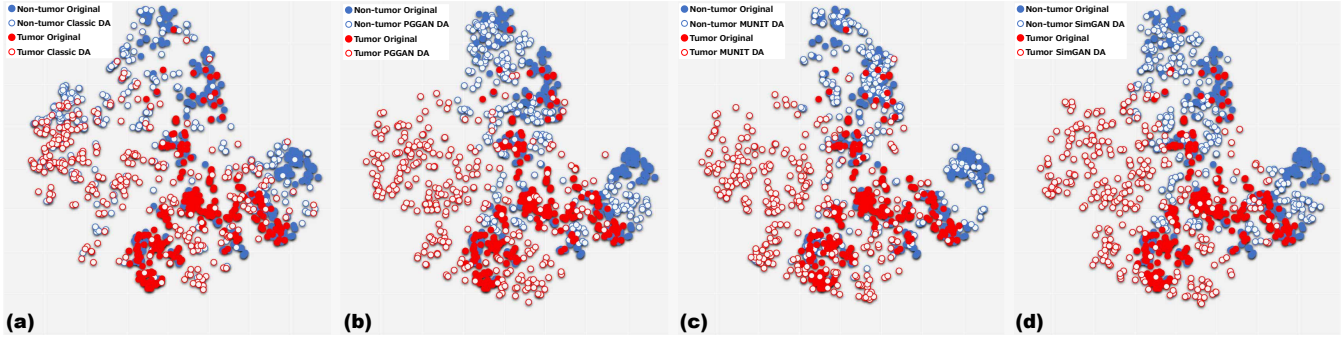

Fig. 10: T-SNE plots with 300 tumor/non-tumor images per each category: Real images *vs* (a) Geometrically-transformed images; (b) PGGAN-generated images; (c) MUNIT-refined images; (d) SimGAN-refined images.

to cover global features and provide the robustness on training for most cases. When combined with the classic DA, the MUNIT-based DA achieves the highest sensitivity 97.48%.

4.3.4 Visual Turing Test Results

Table 3 indicates the confusion matrix for the Visual Turing Test. The expert physician classifies a few PGGAN-generated images as real despite high resolution (i.e., 224×224 pixels).

4.3.5 T-SNE Results

As Fig. 10 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 widely overlap. All GAN-based synthetic images by PGGANs, MUNIT, and SimGAN distribute widely, while their tumor/non-tumor images overlap much less than the geometrically-transformed ones; the MUNIT-refined images show better tumor/non-tumor discrimination and a more similar distribution to the real ones than the other images.

4.4 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 classification sensitivity—especially after refining them with MUNIT or SimGAN, image-to-image GANs. Notably, MUNIT remarkably outperforms SimGAN in terms of sensitivity.

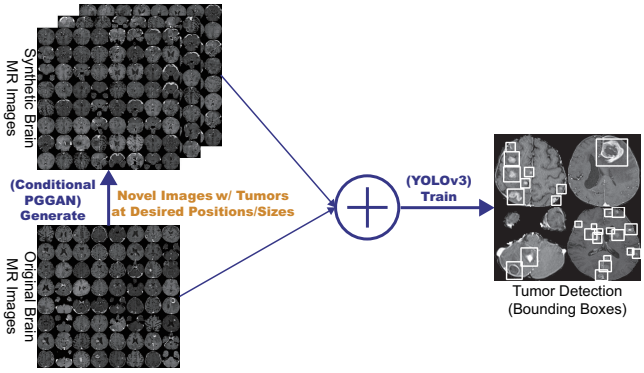

Fig. 11: CPGGAN-based DA for better tumor detection: our CPGGANs generates a number of realistic/diverse brain MR images with tumors at desired positions/sizes based on bounding boxes, and the object detector uses them as additional training data.

5. GAN-based Medical Image Augmentation for 2D Detection

5.1 Motivation

How can we achieve high sensitivity in diagnosis using GANs with minimum annotation cost, based on highly-rough and inconsistent bounding boxes? We aim to generate GAN-based realistic and diverse 256×256 brain MR images with brain metastases at desired positions/sizes for accurate CNN-based tumor detection (Fig. 11). Conventional GANs cannot generate realistic 256×256 whole brain MR images conditioned on tumor positions/sizes under limited training data/highly-rough annotation [13]. Such a high-resolution whole image generation approach, not involving ROIs alone, however, could facilitate detection because it provides more image details and most CNN architectures adopt around 256×256 input pixels. Therefore, we propose CPGGANs, incorporating highly-rough bounding box conditions incrementally into PGGANs [15] to naturally place tumors of random shape at desired positions/sizes on MR images.

Contributions. Our main contributions are as follows:

- **Conditional Image Generation:** As the first bounding box-based 256 × 256 whole pathological image generation approach, CPGGANs can generate realistic/diverse images with objects naturally at desired positions/sizes.
- **Misdiagnosis Prevention:** This study allows us to achieve high sensitivity in automatic CAD using small/fragmented medical imaging datasets with minimum annotation efforts based on highly-rough/inconsistent bounding boxes.
- **Brain Metastases Detection:** This first bounding box-based brain metastases detection method successfully detects tumors with CPGGAN-based DA.

5.2 Materials and Methods

5.2.1 Brain Metastases Dataset

This project uses a dataset of T1c brain axial MR images, collected by the authors: it contains 180 brain metastatic cancer cases from multiple MRI scanners. We also use additional brain MR images from 193 normal subjects only for CPGGAN training, not in tumor detection, to confirm the effect of combining the normal and pathological images for training.

5.2.2 CPGGAN-based Image Generation

Data Preparation For tumor detection, our whole brain metastases dataset (180 patients) is divided into: (*i*) a training set (126 patients); (*ii*) a validation set (18 patients); (*iii*) a test set (36 patients); only the training set is used for GAN training to be fair. Our experimental dataset consists of:

- Training set (2, 813 images/5, 963 bounding boxes);
- Validation set (337 images/616 bounding boxes);
- Test set (947 images/3, 094 bounding boxes).

To confirm the effect of realism and diversity—provided by combining PGGANs and bounding box conditioning—on tumor detection, we compare the following GANs: (*i*) CPGGANs trained only with the brain metastases images; (*ii*) CPGGANs trained also with additional 16, 962 brain images from 193 normal subjects; (*iii*) Image-to-image GAN trained only with the brain metastases images. All images are resized to 256 × 256 pixels (i.e., a power of 2 for better GAN training). As Fig. 12 shows, we lazily annotate tumors with highly-rough and inconsistent bounding boxes to minimize expert physicians' labor.

CPGGANs is a novel conditional noise-to-image training method for GANs, incorporating highly-rough bounding box conditions incrementally into PGGANs [15]. As Fig. 13 shows, we further condition the generator and discriminator to generate realistic and diverse 256 × 256 brain MR images with tumors of random shape at desired positions/sizes using only bounding boxes. Our modifications to the original PGGANs are as follows:

- Conditioning image: prepare a 256 × 256 black image (i.e., pixel value: 0) with white bounding boxes (i.e., pixel value: 255) describing tumor positions/sizes for attention;
- Generator input: resize the conditioning image to the previous generator's output resolution/channel size and concatenate them (noise samples generate the first 4 × 4 images);
- Discriminator input: concatenate the conditioning image with a real or synthetic image.

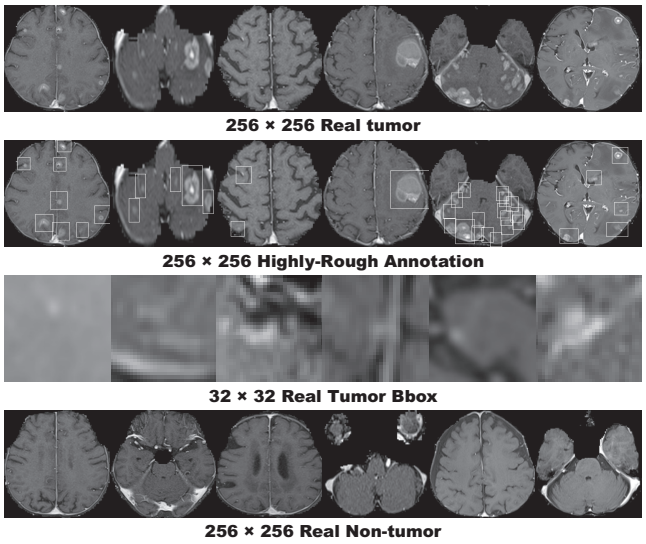

Fig. 12: Example real 256 × 256 MR images with rough annotation used for GAN training and resized 32 × 32 tumor bounding boxes.

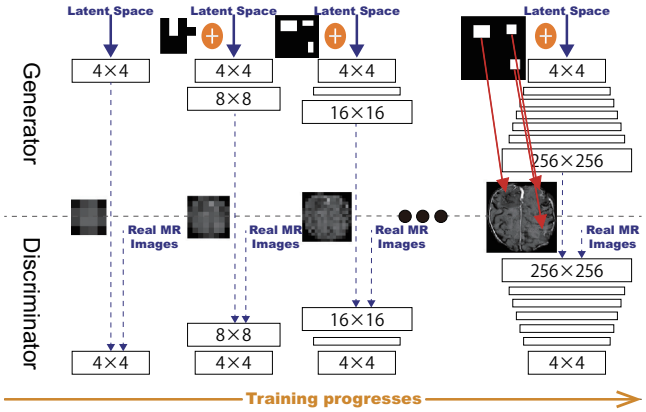

Fig. 13: Proposed CPGGAN architecture for bounding box-based 256×256 MR image generation with tumors at desired positions/sizes.

CPGGAN Implementation Details We use the CPGGAN architecture with the WGAN-GP loss [41]. Training lasts for 3, 000, 000 steps with a batch size of 4 and 2.0×10^{-4} learning rate for the Adam optimizer [42]. During testing, as tumor attention images, we use the annotation of training images with a random combination of horizontal/vertical flipping, width/height shift up to 10%, and zooming up to 10%; these CPGGAN-generated images are used as additional training images for tumor detection.

Image-to-image GAN is a conventional conditional GAN with a U-Net-like [49] generator generating brain tumor images.

5.2.3 YOLOv3-based Brain Metastases Detection

You Only Look Once v3 (YOLOv3) [50] is a fast/accurate CNN-based object detector.

To confirm the effect of GAN-based DA, the following detection results are compared: (*i*) 2, 813 real images without DA, (*ii*), (*iii*), (*iv*) with 4, 000/8, 000/12, 000 CPGGAN-based DA, (*v*), (*vi*), (*vii*) with 4, 000/8, 000/12, 000 CPGGAN-based DA, trained with additional normal brain images, (*viii*), (*ix*), (*x*) with 4, 000/8, 000/12, 000 image-to-image GAN-based DA. Due to the risk of overlooking the diagnosis *via* medical imaging, higher sensitivity matters more than less FPs; thus, we aim to achieve higher sensitivity with acceptable FPs, adding the synthetic training images. Since our annotation is highly-rough, we calculate sensitivity/FPs per slice with both IoU threshold 0.5 and 0.25.

Table 4: Bounding box-based YOLOv3 brain metastases detection results of ten DA setups (with detection threshold 0.1%).

	IoU ≥ 0.5		IoU ≥ 0.25	
	Sensitivity (%)	**FPs per slice**	**Sensitivity (%)**	**FPs per slice**
2,813 real images	67	4.11	83	3.59
+ 4,000 CPGGAN-based DA	**77**	7.64	**91**	7.18
+ 8,000 CPGGAN-based DA	71	6.36	87	5.85
+ 12,000 CPGGAN-based DA	76	11.77	**91**	11.29
+ 4,000 CPGGAN-based DA (+ normal)	69	7.16	86	6.60
+ 8,000 CPGGAN-based DA (+ normal)	73	8.10	89	7.59
+ 12,000 CPGGAN-based DA (+ normal)	74	9.42	89	8.95
+ 4,000 Image-to-Image GAN-based DA	72	6.21	87	5.70
+ 8,000 Image-to-Image GAN-based DA	68	**3.50**	84	**2.99**
+ 12,000 Image-to-Image GAN-based DA	74	7.20	89	6.72

YOLOv3 Implementation Details We use the YOLOv3 architecture with Darknet-53 as a backbone classifier and sum squared error as a loss. During training, we use a batch size of 64 and 1.0×10^{-3} learning rate for the Adam optimizer. The network resolution is set to 416×416 pixels during training and 608×608 pixels during validation/testing, respectively. As classic DA, geometric/intensity transformations are also applied to both real/synthetic images during training to achieve the best performance. For testing, we pick the model with the best sensitivity on validation with detection threshold 0.1%/IoU threshold 0.5 between $96,000$-$240,000$ steps to avoid severe FPs.

5.3 Results

5.3.1 MR Images Generated by CPGGANs

CPGGANs successfully captures the T1c-specific texture and tumor appearance at desired positions/sizes (Fig. 14). Since we use highly-rough bounding boxes, the synthetic tumor shape largely varies within the boxes. When trained with additional normal brain images, it clearly maintains the realism of the original images with less odd artifacts, including tumor bounding boxes, which the additional images do not include.

5.3.2 Brain Metastases Detection Results

Table 4 shows the tumor detection results with/without GAN-based DA. As expected, the sensitivity remarkably increases with the additional synthetic training data while FPs per slice also increase. Surprisingly, adding only $4,000$ CPGGAN-generated images achieves the best sensitivity improvement by 0.10 with IoU threshold 0.5 and by 0.08 with IoU threshold 0.25, due to the real/synthetic training image balance. Fig. 15 also visually indicates that it can alleviate the risk of overlooking the tumor diagnosis with clinically acceptable FPs. Moreover, our results reveal that further realism—associated with the additional normal brain images during training—does not contribute to detection performance, possibly as the training focuses less on tumor generation.

5.4 Conclusion

Our CPGGANs can generate realistic and diverse 256×256 MR images with brain metastases of random shape, unlike rigorous segmentation, naturally at desired positions/sizes, and achieve high sensitivity in tumor detection—even with small/fragmented training data from multiple MRI scanners and lazy annotation using highly-rough bounding boxes; this attributes to the CPGGANs' good generalization ability to incrementally synthesize conditional whole images.

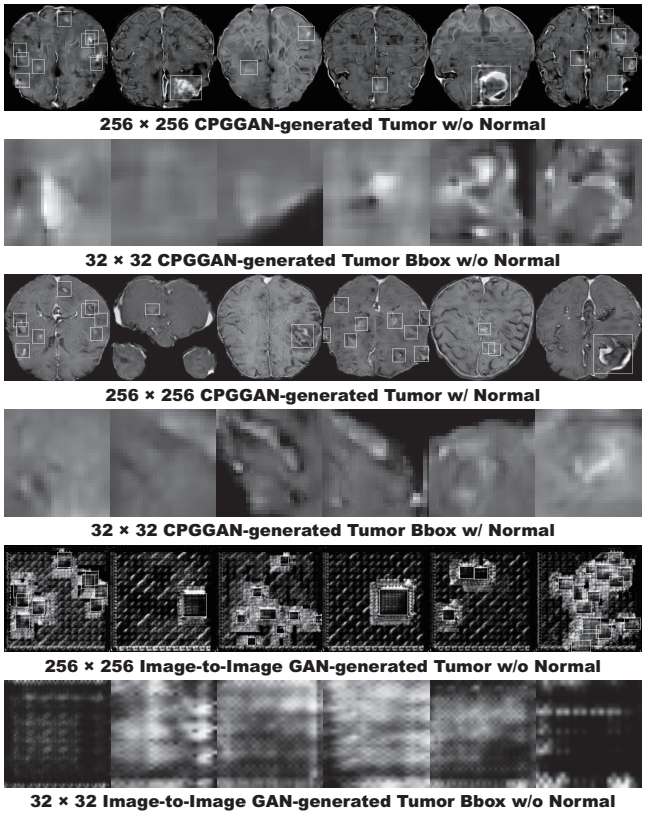

Fig. 14: Example synthetic 256×256 MR images and resized 32×32 tumor bounding boxes yielded by (a), (b) CPGGANs trained without/with additional normal brain images; (c) image-to-image GAN trained without normal images.

6. GAN-based Medical Image Augmentation for 3D Detection

6.1 Motivation

How can GAN generate realistic/diverse 3D nodules placed naturally on lung CT with multiple conditions to boost sensitivity in any 3D object detector? For accurate 3D CNN-based nodule detection (Fig. 16), we propose 3D MCGAN to generate $32 \times 32 \times 32$ nodules. Since nodules vary in position/size/attenuation, to improve CNN's robustness, we adopt two discriminators with different loss functions for conditioning: the context discriminator learns to classify real *vs* synthetic nodule/surrounding pairs with noise box-centered surroundings; the nodule discriminator attempts to classify real *vs* synthetic nodules with size and attenuation conditions.

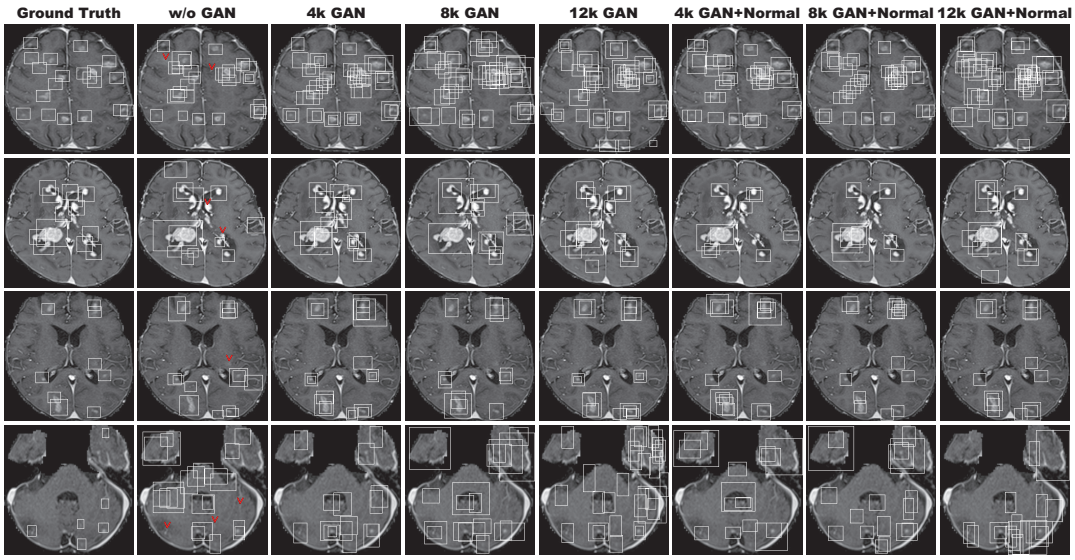

Fig. 15: Example detection results of seven DA setups on four images, compared against the ground truth: (a) ground truth; (b) without CPGGAN-based DA; (c), (d), (e) with 4k/8k/12k CPGGAN-based DA; (f), (g), (h) with 4k/8k/12k CPGGAN-based DA, trained with additional normal brain images. Red V symbols indicate the brain metastases undetected without CPGGAN-based DA, but detected with 4k CPGGAN-based DA.

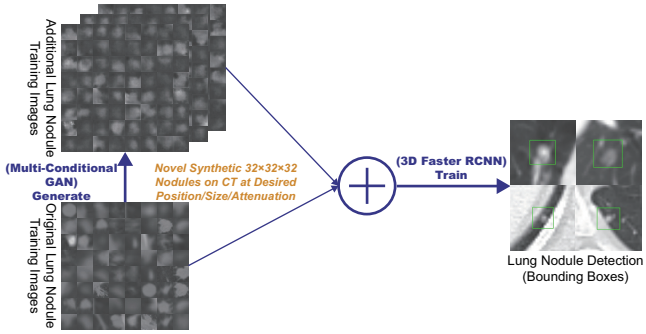

Fig. 16: 3D MCGAN-based DA for better nodule detection: Our MC-GAN generates realistic and diverse nodules naturally on lung CT scans at desired position, size, and attenuation based on bounding boxes, and the CNN-based object detector uses them as additional training data.

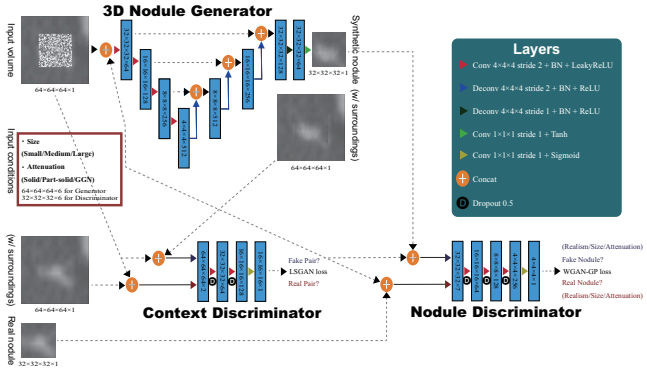

Fig. 17: Proposed 3D MCGAN architecture for realistic/diverse $32 \times 32 \times 32$ lung CT scan of nodule generation: the context discriminator learns to classify real *vs* synthetic nodule/surrounding pairs while the nodule discriminator learns to classify real *vs* synthetic nodules.

Contributions. Our main contributions are as follows:

- **3D Multi-conditional Image Generation:** This first multi-conditional pathological image generation approach shows that 3D MCGAN can generate realistic and diverse nodules placed on lung CT at desired position/size/attenuation.
- **Misdiagnosis Prevention:** This first GAN-based DA method available for any 3D object detector allows to boost sensitivity at fixed FP rates in CAD with limited data.
- **Medical GAN-based DA:** This study implies that training GANs without ℓ_1 loss and using proper augmentation ratio (i.e., $1:1$) may boost CNN-based detection performance with higher sensitivity and less FPs in medical imaging.

6.2 Materials and Methods

6.2.1 3D MCGAN-based Image Generation

Data Preparation This study exploits the Lung Image Database Consortium image collection (LIDC) dataset [51] containing $1,018$ chest CT scans with lung nodules. We only use scans with the slice thickness ≤ 3 mm and 0.5 mm $\leq$ in-plane pixel spacing ≤ 0.9 mm. Then, we interpolate the slice thickness to 1.0 mm and exclude scans with slice number > 400.

To explicitly provide MCGAN with meaningful nodule appearance information and thus boost DA performance, the authors further annotate those nodules by size and attenuation for GAN training with multiple conditions: small (slice thickness ≤ 10 mm); medium (10 mm $\leq$ slice thickness ≤ 20 mm); large (slice thickness > 20 mm); solid; part-solid; Ground-Glass Nodule (GGN). Afterwards, the remaining dataset (745 scans) is divided into: (*i*) a training set (632 scans/$3,727$ nodules); (*ii*) a validation set (37 scans/143 nodules); (*iii*) a test set (76 scans/265 nodules); only the training set is used for MCGAN training to be methodologically sound.

3D MCGAN is a novel GAN training method for DA, generating realistic but new nodules at desired position/size/attenuation, naturally blending with surrounding tissues (Fig. 17). We crop/resize various nodules to $32 \times 32 \times 32$ voxels and replace them with noise boxes from a uniform distribution between $[-0.5, 0.5]$, while maintaining their $64 \times 64 \times 64$ surroundings as VOIs—using those noise boxes, instead of boxes filled with the same voxel values, improves the training robustness; then, we concatenate the VOIs with 6 size/attenuation conditions tiled to $64 \times 64 \times 64$ voxels. So, our generator uses the $64 \times 64 \times 64 \times 7$ inputs to generate desired nodules in the noise box regions. The 3D U-Net [52]-like generator adopts 4 convolutional layers in encoders and 4 decon-

volutional layers in decoders respectively with skip connections to effectively capture both nodule/context information.

We adopt two *Pix2Pix* GAN [53]-like discriminators with different loss functions: the context discriminator learns to classify real *vs* synthetic nodule/surrounding pairs with noise box-centered surroundings using Least Squares loss (LSGANs) [54]; the nodule discriminator attempts to classify real *vs* synthetic nodules with size/attenuation conditions using WGAN-GP [41]. The LSGANs in the context discriminator forces the model to learn surrounding tissue background by reacting more sensitively to every pixel in images than regular GANs. The WGAN-GP in the nodule discriminator allows the model to generate realistic/diverse nodules without focusing too much on details. Empirically, we confirm that such multiple discriminators with the mutually complementary loss functions, along with size/attenuation conditioning, help generate realistic/diverse nodules naturally placed at desired positions on CT scans. We apply dropout to inject randomness and balance the generator/discriminators. Batch normalization is applied to both convolution (using LeakyReLU) and deconvolution (using ReLU).

To confirm the ℓ_1 loss' influence during classifier training, we compare our MCGAN objective without/with it:

$$G^* = \arg\min_G \max_{D1,D2} \mathcal{L}_{\textbf{LSGANs}}(G, D1) + \mathcal{L}_{\textbf{WGAN-GP}}(G, D2), \quad (1)$$

$$G^* = \arg\min_G \max_{D1,D2} \mathcal{L}_{\textbf{LSGANs}}(G, D1) + \mathcal{L}_{\textbf{WGAN-GP}}(G, D2) + 100\mathcal{L}_{\ell_1}(G). \quad (2)$$

3D MCGAN Implementation Details Training lasts for $6,000,000$ steps with a batch size of 16 and 2.0×10^{-4} learning rate for the Adam optimizer. We use horizontal/vertical flipping as DA and flip real/synthetic labels once in three times for robustness. During testing, we augment nodules with the same size/attenuation conditions by applying a random combination to real nodules of width/height/depth shift up to 10% and zooming up to 10% for better DA. We resample the resulting nodules to their original resolution and map back onto the original CT scans to prepare additional training data.

6.2.2 3D Faster RCNN-based Lung Nodule Detection

3D Faster RCNN is a 3D version of Faster RCNN [55] using multi-task loss. To confirm the effect of MCGAN-based DA, we compare the following detection results trained on (*i*) 632 real images without GAN-based DA, (*ii*), (*iii*), (*iv*) with 1×/2×/3× MCGAN-based DA (i.e., 632/1,264/1,896 additional synthetic training images) , (*v*), (*vi*), (*vii*) with 1×/2×/3× MCGAN-based DA trained with ℓ_1 loss. We evaluate the detection performance as follows: Competition Performance Metric (CPM) score [56], average sensitivity at seven pre-defined FP rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan—this quantifies the ability to identify nodules with both very few FPs and moderate FPs.

3D Faster RCNN Implementation Details During training, we use a batch size of 2 and 1.0×10^{-3} learning rate (1.0×10^{-4} after $20,000$ steps) for the SGD optimizer. The input volume size is set to $160 \times 176 \times 224$ voxels. As classical DA, a random combination of width/height/depth shift up to 15% and zooming up to 15% are also applied to both real/synthetic images to achieve the best performance. For testing, we pick the model with the highest sensitivity on validation between $30,000$-$40,000$ steps under IoU threshold 0.25/detection threshold 0.5 to avoid severe FPs.

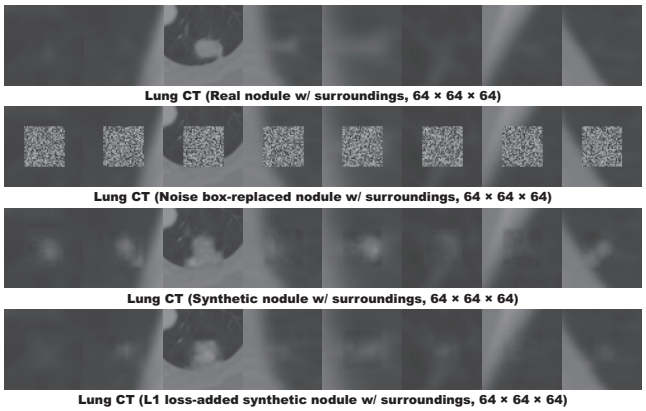

Fig. 18: 2D axial view of example real/synthetic $64 \times 64 \times 64$ CT scans of lung nodules with surrounding tissues; 3D MCGANs generate only $32 \times 32 \times 32$ nodules.

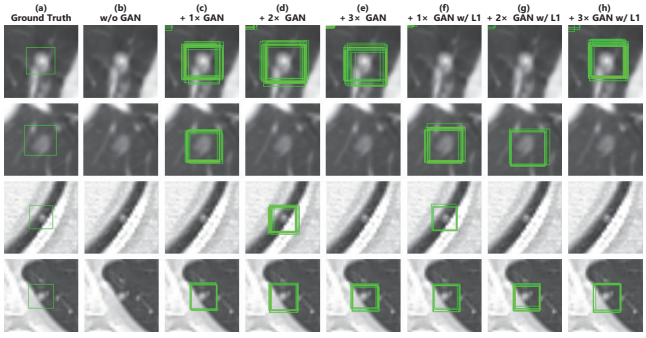

Fig. 19: Example detection results of seven DA setups on four different images, compared against the ground truth (detection threshold 0.5): (a) ground truth; (b) without GAN-based DA; (c), (d), (e) with 1×/2×/3× 3D MCGAN-based DA; (f), (g), (h) with 1×/2×/3× ℓ_1 loss-added 3D MCGAN-based DA.

6.3 Results

6.3.1 Lung Nodules Generated by 3D MCGAN

We generate realistic nodules in noise box regions at various position/size/attenuation, naturally blending with surrounding tissues including vessels, soft tissues, and thoracic walls (Fig. 18). Especially, when trained without ℓ_1 loss, those synthetic nodules look clearly more different from the original real ones.

6.3.2 Lung Nodule Detection Results

Table 5 shows that it is easier to detect nodules with larger size/lower attenuation due to their clear appearance. 3D MCGAN-based DA with less augmentation ratio consistently increases sensitivity at fixed FP rates—especially, training with 1× MCGAN-based DA without ℓ_1 loss outperforms training only with real images under any size/attenuation in terms of CPM, achieving average CPM improvement by 0.032. Fig. 19 visually reveals its ability to alleviate the risk of overlooking the nodule diagnosis with clinically acceptable FPs. Surprisingly, adding more synthetic images tends to decrease sensitivity, due to the real/synthetic training image balance. Moreover, further nodule realism introduced by ℓ_1 loss rather decreases sensitivity as ℓ_1 loss sacrifices diversity in return for the realism.

Table 5: 3D Faster RCNN nodule detection results (CPM) of seven DA setups (IoU ≥ 0.25). Both results without/with ℓ_1 loss at different augmentation ratio are compared. CPM is average sensitivity at $1/8, 1/4, 1/2, 1, 2, 4$, and 8 FPs per scan.

		CPM by Size (%)			CPM by Attenuation (%)		
	CPM (%)	**Small**	**Medium**	**Large**	**Solid**	**Part-solid**	**GGN**
632 real images	51.8	44.7	61.8	62.4	65.5	46.4	24.2
+ 1× 3D MCGAN-based DA	**55.0**	**45.2**	**68.3**	**66.2**	**69.9**	52.1	24.4
+ 2× 3D MCGAN-based DA	52.7	44.7	67.4	42.9	65.5	40.7	**28.9**
+ 3× 3D MCGAN-based DA	51.2	41.1	64.4	66.2	61.6	**57.9**	27.7
+ 1× 3D MCGAN-based DA w/ ℓ_1	50.8	43.0	63.3	55.6	62.6	47.1	27.1
+ 2× 3D MCGAN-based DA w/ ℓ_1	50.9	40.6	64.4	65.4	64.9	43.6	23.3
+ 3× 3D MCGAN-based DA w/ ℓ_1	47.9	38.9	59.4	61.7	59.6	50.7	22.6

6.4 Conclusion

Our bounding box-based 3D MCGAN can generate diverse CT-realistic nodules at desired position/size/attenuation, naturally blending with surrounding tissues—those synthetic training data boost sensitivity under any size/attenuation at fixed FP rates in 3D CNN-based nodule detection. This attributes to the MCGAN's good generalization ability coming from multiple discriminators with mutually complementary loss functions, along with informative size/attenuation conditioning.

7. Discussions on Developing Clinically Relevant AI-Powered Diagnosis Systems

7.1 Feedback from Physicians

7.1.1 Methods for Questionnaire Evaluation

To confirm the clinical relevance for diagnosis of our proposed pathology-aware GAN methods for DA and physician training respectively, we conduct a questionnaire survey for 9 Japanese physicians who interpret MR and CT images in daily practice. The experimental settings are the following:

- **Subjects:** 3 physicians (i.e., a radiologist, a psychiatrist, and a physiatrist) committed to (at least one of) our pathology-aware GAN projects and 6 project non-related radiologists without much AI background.
- **Experiments:** Physicians are asked to answer a questionnaire within 2 weeks from December 6th, 2019 after reading 10 summary slides written in Japanese[*1] about Medical Image Analysis and our pathology-aware GAN projects along with example synthesized images. We conduct both qualitative (i.e., free comments) and quantitative (i.e., five-point Likert scale [57]) evaluation: Likert scale 1 = very negative, 2 = negative, 3 = neutral, 4 = positive, 5 = very positive.

7.1.2 Results

We show the questions and Japanese physicians' response summaries. Concerning the following **Questions 1,2,3**, Fig. 20 visually summarizes the expectation scores on medical AI (i.e., general medical AI, GANs for DA, and GANs for physician training) from both 3 project-related physicians and 6 project non-related radiologists.

Question 1: Are you keen to exploit medical AI in general when it achieves accurate and reliable performance in the near future?

- **Response summary**: As expected, the project-related physicians are AI-enthusiastic while the project non-related radiologists are also generally very positive about the medical AI. Many of them appeal the necessity of AI-based diagnosis for more reliable diagnosis because of the lack of physicians. Meanwhile, other physicians worry about its cost and reliability. We may be able to persuade them by showing expected profitability (e.g., currently CT scanners have an earning rate 16% and CT scans require 2-20 minutes for interpretation in Japan). Similarly, we can explain how experts annotate medical images and AI diagnoses disease based on them (e.g., multiple physicians, not a single one, can annotate the images *via* discussion).

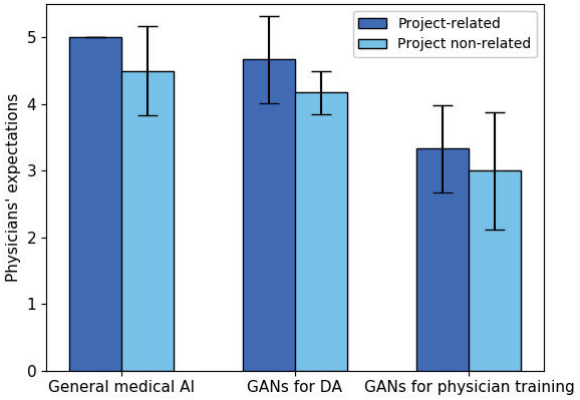

Fig. 20: Bar chart of the expectations on medical AI from 3 project-related physicians and 6 project non-related radiologists, respectively. The vertical rectangles and error bars denote the average five-point Likert scale scores with 95% confidence intervals.

Question 2: What do you think about using GAN-generated images for DA?

- **Response summary**: As expected, the project-related physicians are very positive about the GAN-based DA while the project non-related radiologists are also positive. Many of them are satisfied with its achieved accuracy/sensitivity improvement when available annotated images is limited. However, similarly to their opinions on general Medical Image Analysis, some physicians question its reliability.

Question 3: What do you think about using GAN-generated images for physician training?

- **Response summary**: We generally receive neutral feedback because we do not provide a concrete physician training tool, but instead general pathology-aware generation ideas with example synthesized images—thus, some physicians are positive, and some are not. A physician provides a key idea about a pathology-coverage rate for medical student/expert physician training, respectively. For extensive physician training by GAN-generated atypical images, along with pathology-aware GAN-based extrapolation, further GAN-based extrapolation would be valuable.

Question 4: Any comments/suggestions about our projects towards developing clinically relevant AI-powered systems based on your experience?

[*1] Available *via* Dropbox: `https://www.dropbox.com/sh/bacowc3ilz1p1r3/AABNS9SyjArHq8BntgaODLb2a?dl=0`

- **Response summary**: Most physicians look excited about our pathology-aware GAN-based image augmentation projects and express their clinically relevant requests. The next steps lie in performing further GAN-based extrapolation, developing reliable and clinician-friendly systems with new practice guidelines, and overcoming legal/financial constraints.

7.2 AI and Healthcare Workshop

7.2.1 Methods for Workshop Evaluation

AI and Healthcare sides have a huge gap around technology, funding, and people, such as clinical significance/interpretation, data acquisition, commercial purpose, and anxiety about AI. Aiming to identify/bridge the gap between AI and Healthcare sides in Japan towards develop medical AI fitting into a clinical environment in five years, we hold a workshop for 7 Japanese professionals with various AI and/or Healthcare background. The experimental settings are the following:

- **Subjects:** 2 Medical Imaging experts (i.e., a Medical Imaging researcher and a medical AI startup entrepreneur), 2 physicians (i.e., a radiologist and a psychiatrist), and 3 generalists between Healthcare and Informatics (i.e., a nurse and researcher in medical information standardization, a general practitioner and researcher in medical communication, and a medical technology manufacturer's owner and researcher in health disparities)
- **Experiments:** During the workshop, we conduct 2 activities: (*Learning*) Know the overview of Medical Image Analysis, including state-of-the-art research, well-known challenges/solutions, and the summary of our pathology-aware GAN projects; (*Thinking*) Find the intrinsic gap and its solutions between AI researchers and Healthcare workers after sharing their common and different thinking/working styles. This workshop was held on March 17th, 2019 at Nakayama Future Factory, Open Studio, The University of Tokyo, Tokyo, Japan.

7.2.2 Results

We show the summary of clinically-relevant findings from this Japanese workshop.

Why: Clinical significance/interpretation

- **Challenges**: We need to clarify which clinical situations actually require AI introduction. Moreover, AI's early diagnosis might not be always beneficial for patients.
- **Solutions**: Due to nearly endless disease types and frequent misdiagnosis coming from physicians' fatigue, we should use it as alert to avoid misdiagnosis [58] (e.g., reliable second opinion), instead of replacing physicians. It should help prevent oversight in diagnostic tests not only with CT and MRI, but also with blood data, chest X-ray, and mammography before taking CT and MRI [59]. It could be also applied to segmentation for radiation therapy [60], neurosurgery navigation [61], and pressure ulcers' echo evaluation. Along with improving the diagnosis, it would also make the physicians' workflow easier, such as by denoising [62]. Patients should decide whether they accept AI-based diagnosis under informed consent.

How: Data acquisition

- **Challenges**: Ethical screening in Japan is exceptionally strict, so acquiring and sharing large-scale medical data/annotation are challenging—it also applies to Europe due to General Data Protection Regulation [63]. Considering the speed of technological advances in AI, adopting it for medical devices is difficult in Japan, unlike in medical AI-ready countries, such as the US, where the ethical screening is relatively loose in return for the responsibility of monitoring system stability. Moreover, whenever diagnostic criteria changes, we need further reviews and software modifications; for example, the Tumor-lymph Node-Metastasis (TNM) classification criteria changed for oropharyngeal cancer in 2018 and for lung cancer in 2017, respectively. Diagnostic equipment/target changes also require large-scale data/annotation acquisition again.
- **Solutions**: For Japan to keep pace, the ethical screening should be adequate to the other leading countries. Currently, overseas research and clinical trials are proceeding much faster, so it seems better to collaborate with overseas companies. Moreover, complete medical checkup, which is extremely costly, is unique in East Asia, so Japan could be superior in individuals' multiple medical data—Japan is the only country, where most workers 40 or older are required to have medical checkups once a year independent of their health conditions by the Industrial Safety and Health Act [64]. To handle changes in diagnostic criteria/equipment and overcome dataset/task dependency, it is necessary to establish a common database creation workflow by regularly entering electronic medical records into the database. For reducing data acquisition/annotation cost, GAN-based DA [19] and domain adaptation would be effective.

How: Commercial deployment

- **Challenges**: Hospitals currently do not have commercial benefits to actually introduce medical AI.
- **Solutions**: For example, it would be possible to build AI-powered hospitals [65] operated with less staff. Medical manufacturers could also standardize data format [66], such as for X-ray, and provide some AI services. Many IT giants like Google are now working on medical AI to collect massive biomedical data [67], so they could help rural areas and developing countries, where physician shortage is severe [68], at relatively low cost.

How: Safety and feeling safe

- **Challenges**: Considering multiple metrics, such as sensitivity and specificity, and dataset/task dependency, accuracy could be unreliable, so ensuring safety is challenging. Moreover, reassuring physicians and patients is important to actually use AI in a clinical environment.
- **Solutions**: We should integrate various clinical data, such as blood test biomarkers and multiomics, with images [59]. Moreover, developing bias-robust technology is important since confounding factors are inevitable [69]. To prevent oversight, prioritizing sensitivity over specificity is essential [70]. We should also devise education for medical AI users, such as result interpretation, to reassure patients [71].

8. Conclusion

8.1 Final Remarks

Inspired by their excellent ability to generate realistic and diverse images, we propose to use noise-to-image GANs for (*i*) Medical DA and (*ii*) physician training [9]. Through information conversion, such applications can relieve the lack of pathological data and their annotation; this is uniquely and intrinsically important in Medical Image Analysis, as CNN generalization becomes unstable on unseen data due to large inter-subject, inter-pathology, and cross-modality variability [72]. Towards clinically relevant implementation for the DA and physician training, we find effective loss functions and training schemes for each of them [13], [14]—the diversity matters more for the DA to sufficiently fill the real image distribution whereas the realism matters more for the physician training not to confuse medical students and radiology trainees.

8.2 Future Work

We believe that the next steps towards GAN-based extrapolation and thus atypical pathological image generation lie in (*i*) generation by parts with coordinate conditions [73], (*ii*) generation with both image and gene expression conditions [74], and (*iii*) transfer learning among different body parts and disease types [75]. Due to biological constraints, human interaction is restricted to part of the surrounding environment. Accordingly, we must reason spatial relationships across the surrounding parts to piece them together. Similarly, since machine performance also depends on computational constraints, it is plausible for a generator to generate partial images using the corresponding spatial coordinate conditions—meanwhile, a discriminator attempts to judge realism across the assembled patches by global coherence, local appearance, and edge-crossing continuity. This approach allowed COnditional COordinate GAN (COCO-GAN) to generate state-of-the-art realistic and seamless full images [73]. Since human anatomy has a much stronger local consistency than various object relationships in natural images, reasoning the body's spatial relationships, like the COCO-GAN, would perform effective extrapolation both for medical DA/physician training.

References

[1] E. J. Hwang, S. Park, K. Jin, J. I. Kim, S. Y. Choi, *et al.*, "Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs," *Clin. Infect. Dis.*, vol. 69, no. 5, pp. 739–747, 2018.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[4] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[5] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning [book reviews]," *IEEE T. Neural Networ.*, vol. 20, no. 3, pp. 542–542, 2009.

[6] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 3630–3638, 2016.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, *et al.*, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014.

[8] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2107–2116, 2017.

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

[10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[11] 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.

[12] C. Han, K. Murao, S. Satoh, and H. Nakayama, "Learning more with less: GAN-based medical image augmentation," *Med. Imaging Tech.*, vol. 37, no. 3, pp. 137–142, 2019.

[13] C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, *et al.*, "Infinite brain MR images: PGGAN-based data augmentation for tumor detection," in *Neural Approaches to Dynamics of Signal Exchanges*, Smart Innovation, Systems and Technologies, pp. 291–303, Springer, 2019.

[14] C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, *et al.*, "Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection," *IEEE Access*, vol. 7, pp. 156966–156977, 2019.

[15] 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:1710.10196v3*, 2018.

[16] 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.

[17] P. Stinis, T. Hagge, A. M. Tartakovsky, and E. Yeung, "Enforcing constraints for interpolation and extrapolation in generative adversarial networks," *J. Comput. Phys.*, vol. 397, p. 108844, 2019.

[18] C. Han, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, *et al.*, "Learning more with less: Conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images," in *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 119–127, 2019.

[19] C. Han, Y. Kitamura, A. Kudo, A. Ichinose, L. Rundo, *et al.*, "Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection," in *Proc. International Conference on 3D Vision (3DV)*, pp. 729–737, 2019.

[20] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[21] H. Emami, M. Dong, S. P. Nejad-Davarani, and C. K. Glide-Hurst, "Generating synthetic CTs from magnetic resonance images using generative adversarial networks," *Med. Phys.*, vol. 45, no. 8, pp. 3627–3636, 2018.

[22] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling GANs for data augmentation in mammogram classification," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pp. 98–106, Springer, 2018.

[23] 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.

[24] T. Malygina, E. Ericheva, and I. Drokin, "Data augmentation with GAN: Improving chest X-ray pathologies prediction on class-imbalanced cases," in *Proc. International Conference on Analysis of Images, Social Networks and Texts (AIST)*, pp. 321–334, 2019.

[25] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest X-ray generation and data augmentation for cardiovascular abnormality classification," in *Proc. Medical Imaging: Image Processing*, vol. 10574, p. 105741M, 2018.

[26] T. Kanayama, Y. Kurose, K. Tanaka, K. Aida, S. Satoh, *et al.*, "Gastric cancer detection from endoscopic images using synthesis by GAN," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 530–538, 2019.

[27] C. Gao, S. Clark, J. Furst, and D. Raicu, "Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection," in *Proc. Medical Imaging: Computer-Aided Diagnosis*, vol. 10950, p. 109501K, 2019.

[28] H. C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Interna-*

tional Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 1–11, 2018.

[29] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 732–740, 2018.

[30] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *Proc. International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 29–41, 2019.

[31] S. Wang, "Competencies and experiential requirements in radiology training," in *Radiology Education*, pp. 55–66, Springer, 2012.

[32] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. R. Soc. Interface*, vol. 15, no. 141, p. 20170387, 2018.

[33] 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 *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 240–244, 2018.

[34] M. Prastawa, E. Bullitt, and G. Gerig, "Simulation of brain tumors in MR images for evaluation of segmentation efficacy," *Med. Image Anal.*, vol. 13, no. 2, pp. 297–311, 2009.

[35] 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.

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2234–2242, 2016.

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

[39] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1312.6114*, 2014.

[40] S. Koley, A. K. Sadhu, P. Mitra, B. Chakraborty, and C. Chakraborty, "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.

[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," *arXiv preprint arXiv:1704.00028*, 2017.

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

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

[44] 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.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Learning Representations (ICLR) arXiv preprint arXiv:1502.03167*, 2015.

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

[49] 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.

[50] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[51] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[52] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432, 2016.

[53] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

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

[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99, 2015.

[56] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, "On combining computer-aided detection systems," *IEEE Trans. Med. Imaging*, vol. 30, no. 2, pp. 215–223, 2011.

[57] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Qual. Prog.*, vol. 40, no. 7, pp. 64–65, 2007.

[58] M. E. Vandenberghe, M. L. J. Scott, P. W. Scorer, M. Söderberg, D. Balcerzak, and C. Barker, "Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer," *Sci. Rep.*, vol. 7, p. 45938, 2017.

[59] X. Li, Y. Wang, and D. Li, "Medical data stream distribution pattern association rule mining algorithm based on density estimation," *IEEE Access*, vol. 7, pp. 141319–141329, 2019.

[60] M. Agn, I. Law, P. M. af Rosenschöld, and K. Van Leemput, "A generative model for segmentation of tumor and organs-at-risk for radiation therapy planning of glioblastoma patients," in *Proc. Medical Imaging: Image Processing*, vol. 9784, p. 97841D, 2016.

[61] K. R. Abi-Aad, B. J. Anderies, M. E. Welz, and B. R. Bendok, "Machine learning as a potential solution for shift during stereotactic brain surgery," *Neurosurgery*, vol. 82, no. 5, pp. E102–E103, 2018.

[62] Q. Yang, P. Yan, Y. Zhang, *et al.*, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[63] J. M. M. Rumbold and B. Pierscionek, "The effect of the general data protection regulation on medical research," *J. Med. Internet Res.*, vol. 19, no. 2, p. e47, 2017.

[64] K. Nawata, A. Matsumoto, R. Kajihara, and M. Kimura, "Evaluation of the distribution and factors affecting blood pressure using medical checkup data in Japan," *Health*, vol. 9, no. 1, pp. 124–137, 2016.

[65] A. Chen, Z. Zhang, Q. Li, W. Jiang, Q. Zheng, *et al.*, "Feasibility study for implementation of the AI-powered Internet+ Primary Care Model (AiPCM) across hospitals and clinics in Gongcheng county, Guangxi, China," *Lancet*, vol. 394, p. S44, 2019.

[66] A. Laplante-Lévesque, H. Abrams, M. Bülow, T. Lunner, J. Nelson, *et al.*, "Hearing device manufacturers call for interoperability and standardization of internet and audiology," *Am. J. Audiol.*, vol. 25, no. 3S, pp. 260–263, 2016.

[67] J. Morley, M. Taddeo, and L. Floridi, "Google Health and the NHS: Overcoming the trust deficit," *Lancet Digit. Health*, vol. 1, no. 8, p. e389, 2019.

[68] G. R. Jankharia, "Commentary-radiology in India: The next decade," *Indian J. Radiol. Imaging*, vol. 18, no. 3, p. 189, 2008.

[69] H. Li, G. Jiang, J. Zhang, W. R., W. Z., *et al.*, "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images," *NeuroImage*, vol. 183, pp. 650–665, 2018.

[70] A. Jain, S. Ratnoo, and D. Kumar, "Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach," in *Proc. International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, pp. 1–8, 2017.

[71] S. A. Wartman and C. D. Combs, "Reimagining medical education in the age of AI," *AMA J. Ethics*, vol. 21, no. 2, pp. 146–152, 2019.

[72] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, *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.

[73] C. H. Lin, C. Chang, Y. Chen, D. Juan, W. Wei, and H. Chen, "COCO-GAN: Generation by parts via conditional coordinating," in *Proc. International Conference on Computer Vision (ICCV)*, pp. 4512–4521, 2019.

[74] Z. Xu, X. Wang, H. Shin, D. Yang, H. Roth, *et al.*, "Correlation via synthesis: End-to-end nodule image generation and radiogenomic map learning based on generative adversarial network," *arXiv preprint arXiv:1907.03728*, 2019.

[75] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.