

Department of Creative Informatics
Graduate School of Information Science and Technology

博士論文

Pathology-Aware Generative Adversarial
Networks for Medical Image Augmentation
(医用画像拡張に向けた、
病変部を意識した敵対的生成ネットワーク)

Ph.D. Thesis of: **Changhee Han** (韓 昌熙)
Supervisor: **Hideki Nakayama** (中山 英樹)

ACADEMIC YEAR 2019-2020

Dedicated with love to my mother, for always believing and supporting;
and to my sister, for her unbounded trust and inspiration.

Acknowledgments

Most of all, I want to thank my mother and sister for raising me to value education and science—I would not be writing this thesis without their life-long dedicated support and disarming warmth. My second deepest appreciation goes to Prof. Hideki Nakayama for taking me under his wing and supporting me to explore creative and informative ideas, just like our department name. He kindly allowed me to challenge many different ideas despite my shortcomings. I am grateful to Research Center for Medical Big Data staff Prof. Fuyuko Kido, Dr. Youichirou Ninomiya, Fumika Tamura, and Prof. Shin'ichi Satoh, for offering me amazing research opportunities and helping me apply for Japanese permanent residency. In particular, I always enjoyed discussions and consequent publications together with you, Prof. Murao Kohei.

Many thanks to all people committed to our GAN projects. To physicians, including Dr. Yusuke Kawata and Dr. Fumiya Uchiyama (National Center for Global Health and Medicine), for their brilliant clinical insights and contributions. Especially, I owe entirely to Dr. Tomoyuki Noguchi (Kyushu Medical Center) for always sparing no effort to protect health and save lives with state-of-the-art technology—he introduced me many competent physicians for evaluating our GAN applications' clinical relevance. Dr. Yujiro Furukawa (Jikei University School of Medicine) has always supported our GAN projects from the very first project, trusting its potential despite repeated urgent requests. Dr. Kazuki Umemoto (Juntendo University School of Medicine) is AI-enthusiastic and very passionate about advancing healthcare.

This thesis would not have been achievable without Ryosuke Araki (Chubu University) and Prof. Hideaki Hayashi (Kyushu University) because we launched and carried out these tough but fruitful projects together for so long. During my internship at FujiFilm with Dr. Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, and Dr. Yuanzhong Li, my best memory was drinking endless free beer.

I appreciate those participated in our questionnaire survey and GCL workshop: Dr. Yoshitaka Shida, Dr. Ryotaro Kamei (National Center for Global Health and

Medicine), Dr. Toshifumi Masuda (Kyushu Medical Center), Dr. Ryoko Egashira, Dr. Yoshiaki Egashira, Dr. Tetsuya Kondo (Saga University), Dr. Takafumi Nemoto (Keio University School of Medicine), Dr. Yuki Kaji, Miwa Aoki Uwadaira, Hajime Takeno (The University of Tokyo), and Kohei Yamamoto (Corpy&Co., Inc.).

My sincere thanks go to Prof. Takeo Igarashi, Prof. Tatsuya Harada, Prof. Kazuhiko Ohe, Prof. Manabu Tsukada, and Prof. Murao Kohei for refereeing my Ph.D. thesis. Including Prof. Tsukada's mentoring, I was privileged to have continuous financial and human support from GCL program of my University by MEXT.

Clinically valuable research requires international/interdisciplinary collaboration. Backed up by the GCL, I achieved this goal thanks to Dr. Leonardo Rundo, real friend who invited me to Università degli Studi di Milano-Bicocca twice, University of Cambridge once, and his hometown in Sicily twice. After working hard together everyday from everywhere, he understands this thesis better than anyone in the world. Special thanks also to Prof. Giancarlo Mauri on his 70th birthday in Milan and Prof. Evis Sala in Cambridge, along with the other (mostly Italian) guys in their labs, including Prof. Daniela Besozzi, Prof. Paolo Cazzaniga, Prof. Marco Nobile, Dr. Andrea Tangherloni, and Simone Spolaor. After visiting 19 Italian cities, *sono quasi Italiano!* Nicolas Y. Kröner was my geek tutor at Technische Universität München.

I wish to thank all my pleasant friends for making my Ph.D. days enjoyable and memorable in these intense years. Cheers to our long friendship, UTDS members, including Marishi Mochida and Kazuki Taniyoshi. Again, congratulations on your happy wedding, Yuki/Mari Inoue and Akito/Hitomi Misawa. Moreover, I shared far too many hilarious memories with KEK members Zoltán Ádám Milacski and Florian Gesser. Kurumi Nagasako drew this thesis' very first figure. Prof. Yumiko Furuichi and Miwako Hayasaka are always cheerful and smiling like Mother Teresa.

"My true title of glory is that I will live forever" Last, but not least, I would like to thank Napoleon Bonaparte who taught me that *nothing is impossible*. With courage and hope, I could conduct this research for both human prosperity and individual happiness. Listening to Beethoven's Symphony No.3 Eroica in the morning, I always try to be an everyday hero with great ambition like you. Your reign will never cease.

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 aiming to present 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

\times 256 whole brain MR images with/without tumors separately. Multimodal UN-supervised Image-to-image Translation further refines the synthetic images’ texture and 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 clinically acceptable additional 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 $32 \times 32 \times 32$ 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.

We confirm our pathology-aware GANs’ clinical relevance for diagnosis *via* two discussions: (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.

Contents

1	Introduction	19
1.1	Aims and Motivations	19
1.2	Thesis Overview	23
2	Background	25
2.1	Medical Image Analysis	25
2.2	Deep Learning	26
2.3	Methods to Address Data Paucity	28
3	Investigated Contexts and Applications	31
3.1	GAN-based Medical DA	31
3.2	GAN-based Physician Training	33
4	GAN-based Medical Image Generation	35
4.1	Prologue to First Project	35
4.1.1	Project Publication	35
4.1.2	Context	35
4.1.3	Contributions	36
4.1.4	Recent Developments	36
4.2	Motivation	37
4.3	Materials and Methods	39

4.3.1	BRATS 2016 Dataset	39
4.3.2	DCGAN/WGAN-based Image Generation	39
4.3.3	Clinical Validation <i>via</i> Visual Turing Test	42
4.4	Results	42
4.4.1	MR Images Generated by DCGAN/WGAN	42
4.4.2	Visual Turing Test Results	44
4.5	Conclusion	44

5 GAN-based Medical Image Augmentation for 2D Classification 47

5.1	Prologue to Second Project	47
5.1.1	Project Publications	47
5.1.2	Context	47
5.1.3	Contributions	48
5.1.4	Recent Developments	48
5.2	Motivation	49
5.3	Materials and Methods	51
5.3.1	BRATS 2016 Dataset	51
5.3.2	PGGAN-based Image Generation	51
5.3.3	MUNIT/SimGAN-based Image Refinement	54
5.3.4	ResNet-50-based Tumor Classification	57
5.3.5	Clinical Validation <i>via</i> Visual Turing Test	60
5.3.6	Visualization <i>via</i> t-SNE	61
5.4	Results	62
5.4.1	MR Images Generated by PGGANs	62
5.4.2	MR Images Refined by MUNIT/SimGAN	62
5.4.3	Tumor Classification Results	63
5.4.4	Visual Turing Test Results	66
5.4.5	T-SNE Results	67

5.5	Conclusion	68
-----	----------------------	----

6 GAN-based Medical Image Augmentation for 2D Detection 71

6.1	Prologue to Third Project	71
6.1.1	Project Publication	71
6.1.2	Context	71
6.1.3	Contributions	72
6.1.4	Recent Developments	72
6.2	Motivation	73
6.3	Materials and Methods	76
6.3.1	Brain Metastases Dataset	76
6.3.2	CPGGAN-based Image Generation	76
6.3.3	YOLOv3-based Brain Metastases Detection	79
6.3.4	Clinical Validation <i>via</i> Visual Turing Test	81
6.3.5	Visualization <i>via</i> t-SNE	81
6.4	Results	82
6.4.1	MR Images Generated by CPGGANs	82
6.4.2	Brain Metastases Detection Results	82
6.4.3	Visual Turing Test Results	85
6.4.4	T-SNE Results	86
6.5	Conclusion	87

7 GAN-based Medical Image Augmentation for 3D Detection 91

7.1	Prologue to Fourth Project	91
7.1.1	Project Publication	91
7.1.2	Context	91

7.1.3	Contributions	92
7.1.4	Recent Developments	92
7.2	Motivation	93
7.3	Materials and Methods	95
7.3.1	3D MCGAN-based Image Generation	95
7.3.2	3D Faster RCNN-based Lung Nodule Detection	98
7.3.3	Clinical Validation <i>via</i> Visual Turing Test	99
7.3.4	Visualization <i>via</i> t-SNE	99
7.4	Results	100
7.4.1	Lung Nodules Generated by 3D MCGAN	100
7.4.2	Lung Nodule Detection Results	101
7.4.3	Visual Turing Test Results	103
7.4.4	T-SNE Results	103
7.5	Conclusion	104

8 Discussions on Developing Clinically

Relevant AI-Powered Diagnosis Systems 107

8.1	Feedback from Physicians	107
8.1.1	Methods for Questionnaire Evaluation	107
8.1.2	Results	108
8.2	AI and Healthcare Workshop	113
8.2.1	Methods for Workshop Evaluation	113
8.2.2	Results	114

9 Conclusion 119

9.1	Final Remarks	119
9.2	Future Work	121

A	Scientific Production	137
A.1	Related Publications/Presentations	137
A.2	Other Publications/Presentations	139
A.3	Record of Awards	142

List of Figures

1-1	Conceptual scheme of this thesis	23
4-1	Potential applications of the proposed GAN-based synthetic brain MR image generation.	38
4-2	Example real $128 \times 128/64 \times 64$ MR images used for GAN training. .	40
4-3	Example synthetic $128 \times 128/64 \times 64$ DCGAN-generated MR images.	43
4-4	Example synthetic $128 \times 128/64 \times 64$ WGAN-generated MR images.	43
5-1	Combining noise-to-image and image-to-image GANs for better tumor classification.	50
5-2	Example real 256×256 MR images used for PGGAN training.	52
5-3	PGGAN architecture for 256×256 brain MR image generation.	53
5-4	Example real 256×256 MR image and its geometrically-transformed images.	59
5-5	Example PGGAN-generated 256×256 MR images.	59
5-6	Example PGGAN-generated 256×256 MR images and their refined versions by MUNIT/SimGAN.	63
5-7	ResNet-50 tumor classification results under different training dataset sizes of 12 DA setups, with ImageNet pre-training.	66
5-8	T-SNE plots with 300 256×256 tumor/non-tumor MR images per each category.	67
6-1	CPGGAN-based DA for better tumor detection.	74

6-2	Example real 256×256 MR images with highly-rough annotation used for GAN training and resized 32×32 tumor bounding boxes.	77
6-3	Proposed CPGGAN architecture for synthetic 256×256 brain MR image generation with tumors at desired positions/sizes based on bounding boxes.	78
6-4	Example CPGGAN-generated 256×256 MR images and resized 32×32 tumor bounding boxes.	83
6-5	Example detection results of seven DA setups on four different images, compared against the ground truth.	84
6-6	T-SNE plot with 500 32×32 resized tumor bounding box images per each category.	86
6-7	T-SNE plot with 500 256×256 images per each category.	87
7-1	3D MCGAN-based DA for better nodule detection.	94
7-2	Proposed 3D MCGAN architecture for realistic/diverse $32 \times 32 \times 32$ lung CT scan of nodule generation.	96
7-3	2D axial view of example real/synthetic $64 \times 64 \times 64$ CT scans of nodules with surrounding tissues.	100
7-4	FROC curves of seven DA setups by average/size/attenuation.	101
7-5	Example detection results of seven DA setups on four different images, compared against the ground truth.	102
7-6	T-SNE plot with 500 $32 \times 32 \times 32$ nodule images per each category.	104

List of Tables

4.1	Visual Turing Test results by a physician for classifying 50 real <i>vs</i> 50 synthetic images.	44
5.1	PGGAN architecture details for the generator/discriminator.	54
5.2	MUNIT architecture details for the generator/discriminator.	56
5.3	SimGAN architecture details for the refiner/discriminator.	57
5.4	ResNet-50 architecture details without/with pre-training.	60
5.5	ResNet-50 tumor classification results of 20 DA setups, with (without) ImageNet pre-training.	64
5.6	McNemar’s test p -values for the pairwise comparison of the ResNet-50 tumor classification results by accuracy, sensitivity, and specificity.	65
5.7	Visual Turing Test results by an expert physician for classifying 100 Real (R) <i>vs</i> 100 Synthetic (S) images and 100 Tumor (T) <i>vs</i> 100 Non-tumor (N) images.	67
6.1	Bounding box-based YOLOv3 brain metastases detection results of ten DA setups.	84
6.2	Visual Turing Test results by three physicians for classifying 50 real <i>vs</i> 50 CPGGAN-generated images.	85
7.1	3D Faster RCNN nodule detection results (CPM) of seven DA setups.	101
7.2	Visual Turing Test results by two physicians for classifying 50 real <i>vs</i> 50 3D MCGAN-generated images.	103

8.1 Workshop program to (i) know the overview of Medical Image Analysis and (ii) find the intrinsic gap and solutions between AI researchers and Healthcare workers. 114

List of Abbreviations

AI	Artificial Intelligence
BRATS	Brain Tumor Image Segmentation Benchmark
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CPGGANs	Conditional Progressive Growing of Generative Adversarial Networks
CPM	Competition Performance Metric
CT	Computed Tomography
DA	Data Augmentation
DCGAN	Deep Convolutional Generative Adversarial Network
FLAIR	FLuid Attenuation Inversion Recovery
FP	False Positive
FROC	Free Receiver Operation Characteristic
GAN	Generative Adversarial Network
HGG	High-Grade Glioma
IoU	Intersection over Union
JS	Jensen-Shannon
LGG	Low-Grade Glioma
LSGANs	Least Squares Generative Adversarial Networks
MCGAN	Multi-Conditional Generative Adversarial Network
MRI	Magnetic Resonance Imaging
MUNIT	Multimodal UNsupervised Image-to-image Translation
PGGANs	Progressive Growing of Generative Adversarial Networks
ReLU	Rectified Linear Unit

ROI	Region Of Interest
SGD	Stochastic Gradient Descent
SimGAN	Simulated and unsupervised learning Generative Adversarial Network
t-SNE	t-distributed Stochastic Neighbor Embedding
T1	T1-weighted
T1c	contrast enhanced T1-weighted
T2	T2-weighted
UNIT	UNsupervised Image-to-image Translation
VAE	Variational AutoEncoder
VOI	Voxel Of Interest
WGAN	Wasserstein Generative Adversarial Network
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty

Chapter 1

Introduction

“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]

1.1 Aims and Motivations

Convolutional Neural Networks (CNNs) have revolutionized Medical Image Analysis by extracting valuable insights for better clinical examination and medical intervention; the CNNs occasionally outperformed even 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 Aug-

mentation (DA) using natural images, such as 21% performance improvement in eye-gaze estimation [8].

Interpolation refers to new data point construction within a discretely-sampled data distribution. In terms of the interpolation, the GAN-based image augmentation is reliable on the medical images because medical modalities (e.g., X-ray, CT, MRI) can display the human body’s strong anatomical consistency at fixed position while clearly reflecting inter-subject variability [9, 10]—this is different from the natural images, where various objects can appear at any position; accordingly, to tackle large inter-subject, inter-pathology, and cross-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 [11]. The noise-to-image GAN training is much more difficult 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 [12]. For the physician training, the GANs can display novel desired pathological images and help train medical trainees despite infrastructural and legal constraints [13]. 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 [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 the medical students and radiology trainees.

So, how can we perform clinically relevant GAN-based DA/physician training using only limited annotated training images? Always in collaboration with physicians, for improving 2D classification, we combine the noise-to-image [15, 16] (i.e., Progressive Growing of GANs, PGGANs [17]) and image-to-image GANs (i.e., Multimodal UNsupervised Image-to-image Translation, MUNIT [18]); the two-step GAN can generate and refine realistic/diverse original-sized 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, respectively. To meet both clinical demands, we propose novel 2D/3D bounding box-based GANs conditioned on pathology position/size/appearance; the bounding box-based detection requires much less physicians’ annotation effort than rigorous 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 [19]; inevitable errors, not limited between two data points, caused by the model reduction forces a generator to synthesize images that the generator 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 [20]. 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 $32 \times 32 \times 32$ lung nodules placed naturally at desired position/size/attenuation on CT scans [21]; 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, we confirm our pathology-aware GANs’ clinical relevance for diagnosis as a clinical decision support system and non-expert physician training tool *via* two

discussions: (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, always 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.

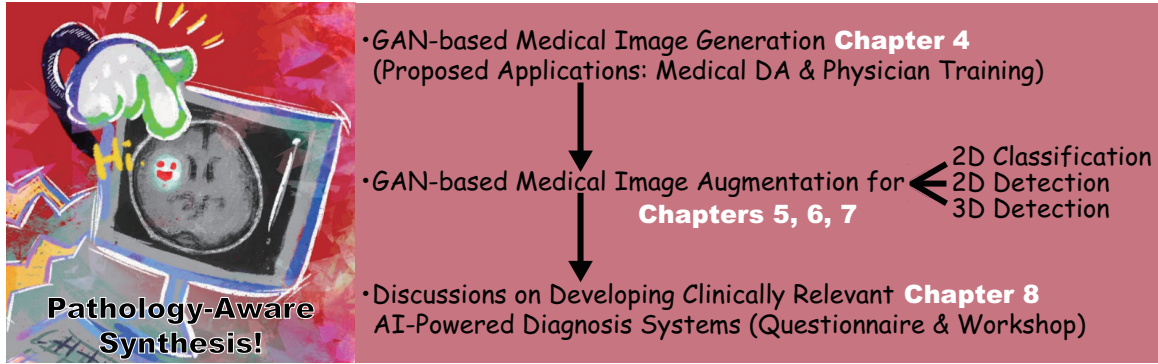


Figure 1-1: Conceptual scheme of this thesis: inspired by their ability to generate realistic/diverse medical images, we propose novel noise-to-image GAN-based clinical applications, (i) medical DA and (ii) physician training; then, to present such GAN applications’ technical soundness, we successfully 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.

1.2 Thesis Overview

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. 1-1). **Chapter 2** covers the background of Medical Image Analysis and Deep Learning, as well as methods to address data paucity to bridge them. **Chapter 3** 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 **Chapter 4**. **Chapter 5** presents a two-step GAN for 2D classification that combines both noise-to-image and image-to-image GANs. **Chapter 6** proposes CPGGANs 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 **Chapter 7**. **Chapter 8** 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, **Chapter 9** provides the conclusive remarks and future directions for further GAN-based extrapolation.

Chapter 2

Background

This chapter introduces basic concepts in Medical Image Analysis and Deep Learning. Afterwards, we describe methods to address data paucity because they play the greatest role in bridging the Medical Image Analysis and Deep Learning.

2.1 Medical Image Analysis

Medical Image Analysis refers to the process of increasing clinical examination/medical intervention efficiency, based on several imaging modalities and digital image processing techniques [22, 23]; to effectively visualize the human body’s anatomical and physiological features, it covers various modalities including X-ray, CT, MRI, positron emission tomography, endoscopy, optical coherence tomography, pathology, ultrasound imaging, and fundus imaging. Its tasks are mainly classified into three groups: (i) Early detection/diagnosis/prognosis of disease often based on pathology classification/detection/segmentation and survival prediction [24, 25]; (ii) Clinical workflow enhancement often based on body part segmentation, inter-modality registration, 3D reconstruction, flow measurement, and surgery simulation [26, 27]; (iii) Clinically impossible image analysis, such as radiogenomics that identifies the correlation between cancer imaging features and gene expression [28].

Among the various modalities, this thesis focuses on the most common 3D modalities for non-invasive diagnosis, CT and MRI. To get a detailed picture inside the body,

the CT merges multiple X-rays at different angles using computational tomographic reconstruction [9, 29]. Since X-ray intensity is associated with the mass attenuation coefficient, higher-density tissues show higher attenuation and *vice versa*. Accordingly, each voxel possesses its attenuation value following the Hounsfield scale from $-1,000$ to $+1,000$ (e.g., Hounsfield units $-1,000$ for air, 0 for water, and $+1,000$ for dense bone). The CT can provide a outstanding contrast within soft-tissue, bone, and lung while the soft-tissue contrast is poor—accordingly, it is especially performed for comprehensive lung assessment.

The MRI uses magnetization properties of atomic nuclei [10, 30]. Since different tissues show various relaxation processes when the nuclei return to their resting alignment, the tissues’ proton density maps serve as both anatomical and functional images. Since the tissues possess two different relaxation times, T1 (i.e., longitudinal relaxation time) and T2 (i.e., transverse relaxation time), as MRI sequences, we can obtain both T1-weighted (T1) and T2-weighted (T2) images. Moreover, using very long repetition time and time to echo, we can obtain FLuid Attenuation Inversion Recovery (FLAIR) images. The MRI can provide a superior soft-tissue contrast to the CT—accordingly, it is especially performed for comprehensive brain assessment.

2.2 Deep Learning

Deep Learning is a kind of Machine Learning algorithms, based on Artificial Neural Networks [31]. The Deep Neural Networks consist of many linearly connected non-linear units whose parameters are optimized by gradient descent [32]; accordingly, their multiple layers can gradually grasp more-detailed features as training progresses (i.e., learning which features to place is automatic). Thanks to the good generalization ability, under large-scale data, the Deep Learning significantly outperforms classical Machine Learning algorithms relying on feature engineering. A visual cortex includes arrangements of simple and complex cells activated by a receptive field (i.e., subregions of a visual field); inspired by this biological structure [33], CNNs adopt a mathematical operation called convolution to achieve translation in-

variance [34]. Since the CNNs are excellent at image/video recognition, their diverse medical applications include pathology classification/detection/segmentation and survival prediction [24, 25].

Variational AutoEncoders (VAEs) often suffer from blurred samples despite easier training, due to the imperfect reconstruction using a single objective function [35]; meanwhile, GANs have revolutionized image generation in terms of realism and diversity [36], including denoising [37] and MRI-to-CT translation [38], based on a two-player objective function using two CNNs [7]: a generator G tries to generate realistic images to fool a discriminator D while maintaining diversity; D attempts to distinguish between the real and synthetic images. However, difficult GAN training from the two-player objective function accompanies artifacts and mode collapse [39], when generating high-resolution images (e.g., 256×256 pixels) [40]; to tackle this, multi-stage noise-to-image GANs have been proposed: AttnGAN generated images from text using attention-based multi-stage refinement [41]; PGGANs generated realistic images using low-to-high resolution multi-stage training [17].

Contrarily, to obtain images with desired texture and shape, some researchers have proposed image-to-image GANs: MUNIT translated images using both GANs and VAEs [18]; Simulated and unsupervised learning GAN (SimGAN) translated images for DA using the self-regularization term and local adversarial loss [8]; Isola *et al.* proposed *Pix2Pix* GAN to produce robust images using paired training samples [42]. Others have proposed conditional GANs: Reed *et al.* proposed bounding box-based conditional GAN to control generated images' local properties [43]; Park *et al.* proposed multi-conditional GAN to refine base images based on texts describing desired position [44].

In Healthcare, medical images have generated the largest volume of data and this trend will no doubt increase due to equipment improvement [45, 46]. Accordingly, as the Deep Learning dominates Computer Vision, Medical Image Analysis is not an exception; their combination can analyze the large-scale medical images and extract valuable insights for better clinical examination and medical intervention. However, the biggest challenge to bridge them lies in the difficulty of obtaining desired patho-

logical images, especially for rare disease [3, 4]. Moreover, it is time-consuming and observer-dependent, even for expert physicians, to annotate them.

2.3 Methods to Address Data Paucity

So, how can we tackle the data paucity? We can either attempt to (a) overcome the lack of generalization or (b) overcome difficulties in optimization. The most straightforward and effective way to address the generalization is DA [47, 48]; because the best model when given data is uncertain, we commonly increase training set size. Human perception is invariant to size, shape, brightness, and color [49]. Accordingly, we recognize the same objects while their such features change, and thus intentionally changing the features is plausible to obtain more data. Such classical DA include (i) x/y/z-axis flipping and rotating, (ii) zooming and scaling, (iii) cropping, (iv) translating, (v) elastic deformation, (vi) adding Gaussian noise (i.e, the distortion of high frequency features), and (vii) brightness and contrast fluctuation.

Recent DA techniques focus on regularization: Mixup [50] and Between-class learning [51] mixed two images during training, such as a dog image and a cat one, for regularization; Cutout randomly masked out square regions during training for regularization [52]; CutMix combined the Mixup and Cutout [53]. As a recent impressive DA approach, AutoAugment automatically searched for improved DA policies [54]. Moreover, similarly to the Mixup among all images within the same class, GAN-based DA can fill the uncovered real image distribution by generating realistic and diverse images *via* many-to-many mapping [55].

Along with the DA, researchers proposed many other techniques to improve the generalization: semi-supervised learning can considerably increase accuracy under limited labeled data by using pseudo labels for unlabeled data [5]; unsupervised anomaly detection allows to detect out-of-distribution images from normal ones, such as disease, without any labeled data [56]; regularization techniques, such as dropout [57], Lasso [58], and elastic net [59], are commonly used for reducing overfitting; similarly, ensembling multiple models *via* bagging [60] and boosting [61] can

effectively increase the robustness; Lastly, in Medical Image Analysis, we can fuse multiple image modalities and/or sequences, such as MRI + CT [62] and T1 MRI + T2 MRI [63].

Moreover, many techniques exist for overcoming the difficulties in optimization: transfer learning can achieve better parameter initialization [64]; problem reduction, such as inputting 2D/3D image patches instead of a whole image, can eliminate unnecessary parameters [65]; learning methods with less data, such as zero-shot learning [66], one-shot learning [6], and neural Turing machine [67], are also promising; meta-learning promotes a versatile model applicable to various tasks without requiring multiple training from scratch [68].

Chapter 3

Investigated Contexts and Applications

In terms of interpolation, GAN-based medical image augmentation is reliable because medical modalities (e.g., X-ray, CT, MRI) can display the human body’s strong anatomical consistency at fixed position while clearly reflecting inter-subject variability [9, 10]—this is different from natural images, where various objects can appear at any position. Accordingly, we proposed to use noise-to-image GANs for (i) medical DA and (ii) physician training [11] in **Chapter 4**. Since then, research towards such clinically valuable applications has shown great promise. This chapter covers such related research works except our own works [15, 16, 20, 21] included in **Chapters 5-7**. Involving 9 physicians, we discuss in detail the clinical relevance of the GAN-based medical DA and physician training in **Chapter 8**.

3.1 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 [37] and MRI-to-CT translation [38]: Wu *et al.* translated 256×256 normal mammograms into lesion ones [69], Gupta *et al.* translated 1024×512 normal leg X-ray images into

bone lesion ones [70], and Malygina *et al.* translated $256 \times 256/512 \times 512$ normal chest X-ray images into pneumonia/pleural-thickening ones [71]. 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 [12], Madani *et al.* augmented 128×128 chest X-ray images with cardiovascular abnormality [72], and Konidaris *et al.* augmented 192×160 brain MR images with Alzheimer’s disease [73].

To facilitate pathology detection and segmentation, researchers conditioned the image-to-image GANs, not the noise-to-image GANs like our work in **Chapter 6**, 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 [19]. To the best of our knowledge, only Kanayama *et al.* tackled bounding box-based pathology detection using the image-to-image GAN [74]; 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 using binary classification [75].

Since 3D imaging is spreading in radiology (e.g., CT and 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 [76],

similarly to the Kanayama *et al.*'s work [74]; 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 [77]. Recently, instead of generating realistic images and training classifiers on them separately, Chaitanya *et al.* directly optimized segmentation results on cardiac MR images [26]; however, it segmented body parts, instead of pathology. Since effective GAN-based medical DA generally requires much engineering effort, we also published a tutorial journal paper [14] about tricks to boost classification/detection/segmentation performance using the GANs, based on our experience and related work.

3.2 GAN-based Physician Training

While medical students and radiology trainees must view thousands of images to become competent [78], accessing such abundant medical images is often challenging due to infrastructural and legal constraints [79]. 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 [80]; 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 [76]; 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 [13].

Chapter 4

GAN-based Medical Image Generation

4.1 Prologue to First Project

4.1.1 Project Publication

- **GAN-based Synthetic Brain MR Image Generation.** C. Han, H. Hayashi, L. Rundo, R. Araki, Y. Furukawa, W. Shimoda, S. Muramatsu, G. Mauri, H. Nakayama, In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 734–738, April 2018.

4.1.2 Context

Prior to this work, it remained challenging to generate realistic and diverse medical images using noise-to-image GANs, not image-to-image GANs [37], due to their unstable training. GAN architectures well-suited for medical images were unclear. Yi *et al.* published results on the noise-to-image GAN-based brain MR image generation, proposing its potential for medical DA and physician training while our paper was under submission [81]; however, they only generated single-sequence low-resolution 128×64 brain MR images without tumors.

4.1.3 Contributions

This project’s main contribution is to propose to use recently developed Wasserstein Generative Adversarial Network (WGAN) [82] for medical DA and physician training—the medical GAN applications are reliable in terms of interpolation because medical modalities can display the human body’s strong anatomical consistency at fixed position while clearly reflecting inter-subject variability. We also demonstrate the noise-to-image GAN’s such potential by generating multi-sequence realistic/diverse 128×128 whole brain tumor MR images [82]; then, we confirm the superb realism *via* Visual Turing Test by a physician.

4.1.4 Recent Developments

Since proposing the GAN applications, we have successfully applied the noise-to-image GANs to improve 2D tumor classification/detection on 256×256 brain MR images [15, 16, 20] as described in **Chapters 5 and 6**. For better 3D tumor segmentation, Shin *et al.* have translated $128 \times 128 \times 54$ normal brain MR images into tumor ones using the image-to-image GAN [76]. Finlayson *et al.* have generated 1024×1024 pelvic fracture/non-fracture X-ray images using a class-conditional noise-to-image GAN, also introducing ongoing work on validating such GANs’ radiology educational efficacy [13]. Kwon *et al.* have generated realistic/diverse 3D brain MR images using the noise-to-image GAN [83].

4.2 Motivation

Along with classic methods [84], CNNs have recently revolutionized medical image analysis [85], including brain MRI segmentation [86]. However, CNN training demands extensive medical data that are laborious to obtain [87]. To overcome this issue, DA techniques via reconstructing original images are common for better performance, such as geometry and intensity transformations [88, 89].

However, those reconstructed images intrinsically resemble the original ones, leading to limited performance improvement in terms of generalization abilities; thus, generating realistic (similar to the real image distribution) but completely new images is essential. In this context, GAN-based DA has excellently performed in general computer vision tasks. It attributes to GAN’s good generalization ability from matching the noise-generated distribution to the real one with a sharp value function. Especially, Shrivastava *et al.* (SimGAN) outperformed the state-of-the-art with a relative 21% improvement in eye-gaze estimation [8].

So, 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. 4-1) [90]. 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) [40] and WGAN [82] to generate realistic images, and an expert physician validates them *via* Visual Turing Test [91].

Research Questions. We mainly address two questions:

- **GAN Selection:** Which GAN architecture is well-suited for realistic medical image generation?
- **Medical Image Processing:** How can we handle MR images with specific intra-sequence variability?

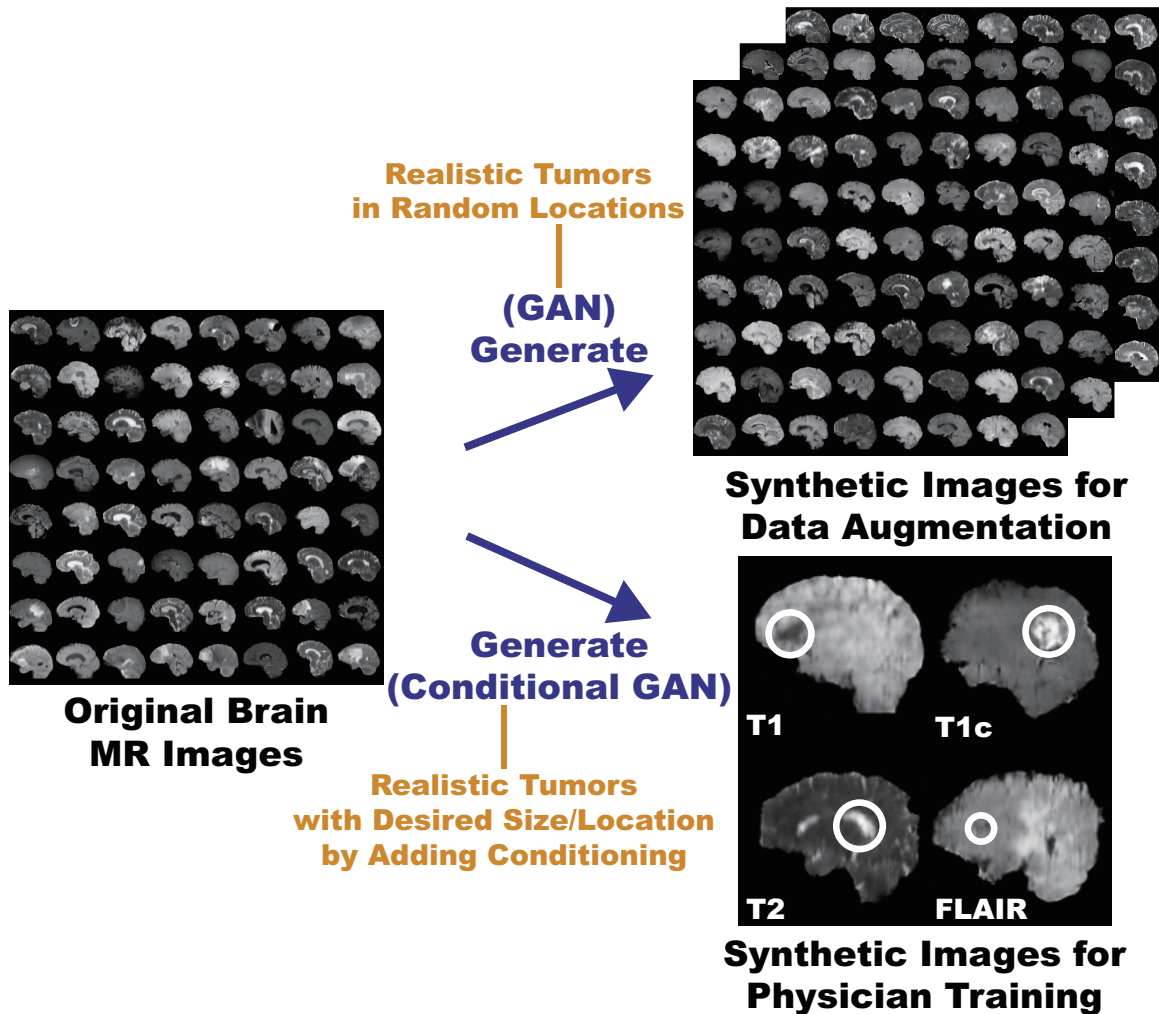


Figure 4-1: Potential applications of the proposed GAN-based synthetic brain MR image generation: (1) DA for better diagnostic accuracy by generating random realistic images giving insights in classification; (2) physician training for better understanding various diseases to prevent misdiagnosis by generating desired realistic pathological images.

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 valuable clinical applications: 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.

4.3 Materials and Methods

Towards clinical applications utilizing realistic brain MR images, we generate synthetic brain MR images from the original samples using GANs. Here, we compare the most used two GANs, DCGAN and WGAN, to find a well-suited GAN between them for medical image generation—it must avoid mode collapse and generate realistic MR images with high resolution.

4.3.1 BRATS 2016 Dataset

This project exploits a dataset of multi-sequence brain MR images to train GANs with sufficient data and resolution, which was originally produced for the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) Challenge [92]. In particular, the BRATS 2016 training dataset 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—they were skull stripped and resampled to isotropic $1\text{mm} \times 1\text{mm} \times 1\text{mm}$ resolution with $240 \times 240 \times 155$ voxels; among the different sectional planes, we use sagittal multi-sequence scans of the HGG patients to show that our GANs can generate a complete view of the whole brain anatomy (allowing for visual consistency among the different brain lobes), including also severe tumors for clinical purpose.

4.3.2 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 (DCGAN architecture results in stable training on 64×64 pixels [40], and so 128×128 is reasonably a high-resolution). Fig. 4-2 shows example real MR images used for training; each sequence contains 15,400 images with $220 \text{ patients} \times 70 \text{ slices}$ (61,600 in total).

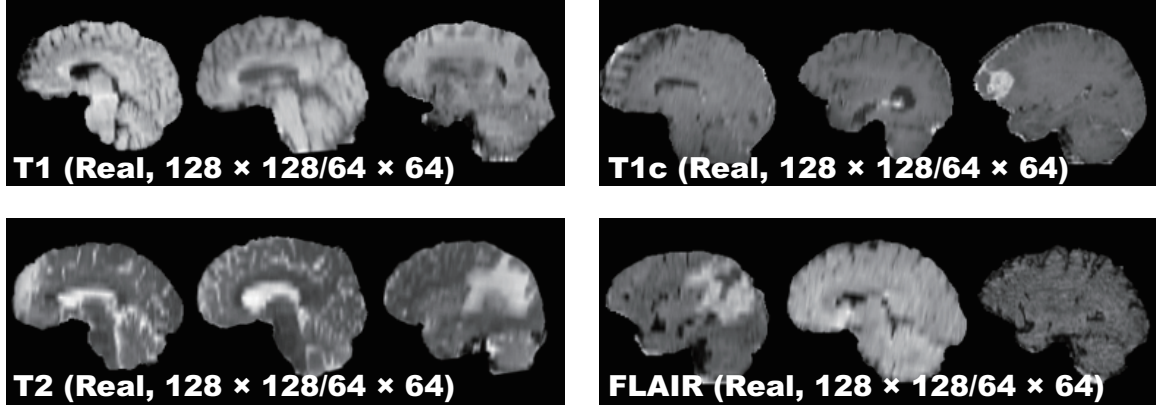


Figure 4-2: Example real $128 \times 128/64 \times 64$ MR images used for GAN training: the resized sagittal multi-sequence brain MRI scans of patients with HGG on the BRATS 2016 training dataset [92].

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. We also generate 64×64 Concat images to compare the generation performance in terms of image resolution.

DCGAN [40] is a standard GAN [7] with a convolutional architecture for unsupervised learning; this generative model uses up-convolutions interleaved with Rectified Linear Unit (ReLU) non-linearity and batch normalization.

Let p_{data} be a generating distribution over data \mathbf{x} . The generator $G(\mathbf{z}; \theta_g)$ is a

mapping to data space that takes a prior on input noise variables $p_z(\mathbf{z})$, where G is a neural network with parameters θ_g . Similarly, the discriminator $D(\mathbf{x}; \theta_d)$ is a neural network with parameters θ_d that takes either real data or synthetic data and outputs a single scalar probability that \mathbf{x} came from the real data. The discriminator D maximizes the probability of classifying both training examples and samples from G correctly while the generator G minimizes the likelihood; it is formulated as a minimax two-player game with value function $V(G, D)$:

$$\begin{aligned} \min_G \max_D V(D, G) = & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] \\ & + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \end{aligned} \quad (4.1)$$

This can be reformulated as the minimization of the Jensen-Shannon (JS) divergence between the distribution p_{data} and another distribution p_g derived from p_z and G .

DCGAN Implementation Details We use the same DCGAN architecture [40] 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 [82] is an alternative to traditional GAN training, as the JS divergence is limited, such as when it is discontinuous; this novel GAN achieves stable learning with less mode collapse by replacing it to the Earth Mover (EM) distance (i.e., the Wasserstein-1 metrics):

$$W(p_g, p_r) = \inf_{p \in \Pi(p_g, p_r)} \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim p} \|\mathbf{x} - \mathbf{x}'\|, \quad (4.2)$$

where $\Pi(p_g, p_r)$ is the set of all joint distributions p whose marginals are p_g and p_r , respectively. In other words, p implies how much mass must be transported from one distribution to another. This distance intuitively indicates the cost of the optimal transport plan.

WGAN Implementation Details We use the same DCGAN architecture [40] for WGAN training. A batch size of 64 and Root Mean Square Propagation (RMSprop) optimizer with 5.0×10^{-5} learning rate were implemented.

4.3.3 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; Concat images were classified together with real T1, T1c, T2, and FLAIR images in equal proportion. The so-called Visual Turing Test [91] uses binary questions to probe a human ability to identify attributes and relationships in images. For these motivations, it is commonly used to evaluate GAN-generated images, such as for SimGAN [8]. This applies also to medical images in clinical environments [80], wherein physicians' expertise is critical.

4.4 Results

This section shows how DCGAN and WGAN generate synthetic brain MR images. The results include instances of synthetic images and their quantitative evaluation of the realism by an expert physician.

4.4.1 MR Images Generated by DCGAN/WGAN

DCGAN Fig. 4-3 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 analogous to mode collapse for 64×64 Concat images, while sharing the same hyper-parameters with 128×128 . **WGAN** Fig. 4-4 shows the example output of WGAN in each sequence. Remarkably outperforming DCGAN, WGAN successfully captures the sequence-specific texture

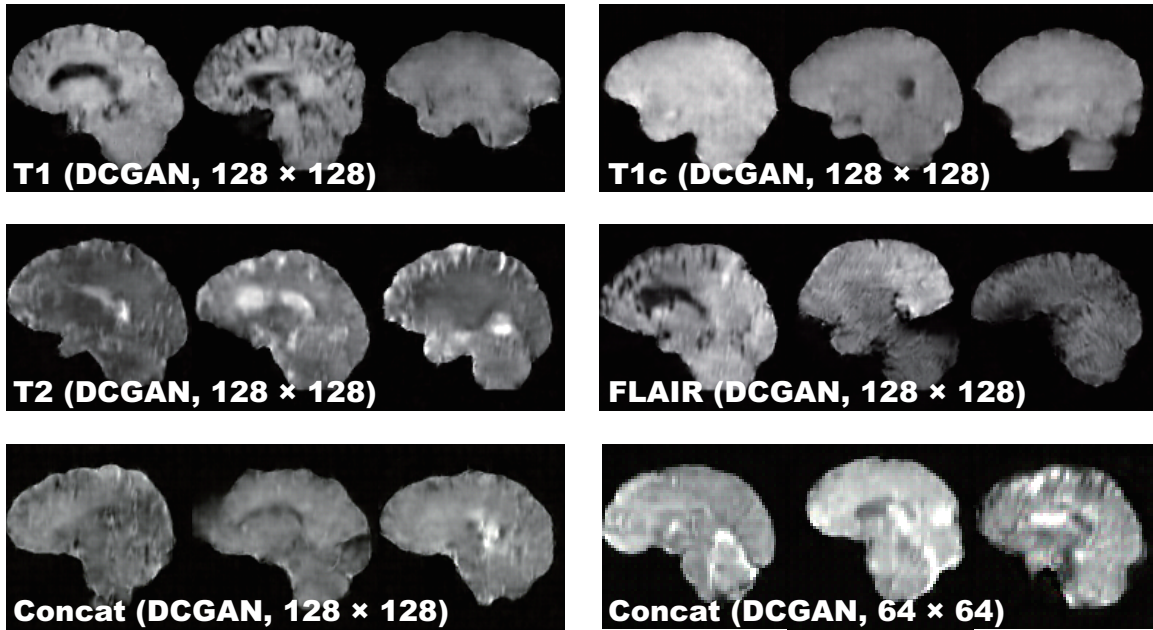


Figure 4-3: Example synthetic $128 \times 128/64 \times 64$ DCGAN-generated MR images.

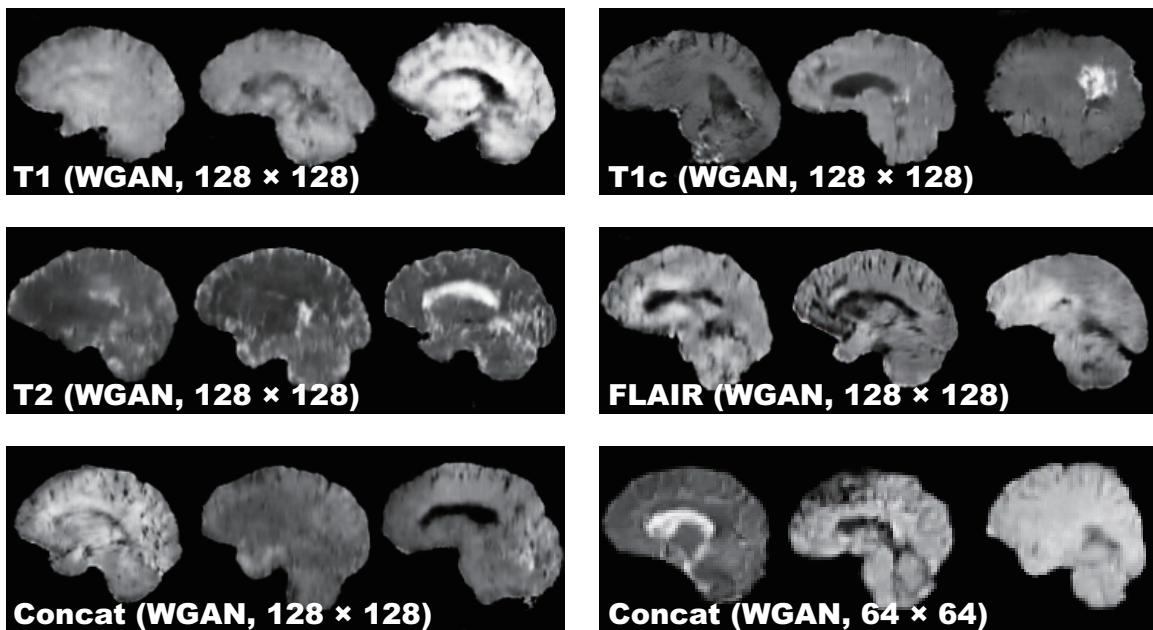


Figure 4-4: Example synthetic $128 \times 128/64 \times 64$ WGAN-generated MR images.

and tumor appearance while maintaining the realism of the original brain MR images. As expected, 128×128 Concat images tend to have more messy and unrealistic artifacts than 64×64 Concat ones, especially around boundaries of the brain, due to the introduction of unexpected intensity patterns.

Table 4.1: 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

4.4.2 Visual Turing Test Results

Table 4.1 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 unfamiliar in clinical routine, even for highly hyper-intense 64×64 Concat images by DCGAN; distinguishing Concat images was easier compared to the case of T1, T1c, T2, and FLAIR images because the physician often felt odd from the artificial sequence. WGAN succeeded to deceive the physician significantly better than DCGAN for all the MRI sequences except FLAIR images (62% to 54%).

4.5 Conclusion

Our preliminary results show that GANs, especially WGAN, can generate 128×128 realistic multi-sequence brain MR images that even an expert physician is unable to accurately distinguish from the real, leading to valuable clinical applications, such as DA and physician training. This attributes to WGAN’s good generalization ability with a sharp value function. In this context, DCGAN might be unsuitable due to both inferior realism and mode collapse in terms of intensity. We only use slices of interest in training to obtain desired MR images and generate both original/Concat sequence images for DA in medical imaging.

This study confirms the synthetic image quality by the human expert evaluation, but a more objective computational evaluation for GANs should also follow, such as Classifier Two-Sample Tests (C2ST) [93], which assesses whether two samples are drawn from the same distribution. Currently this work uses sagittal MR images alone, so we plan to generate coronal and transverse images. As this research uniformly selects middle slices in pre-processing, better data generation demands developing a classifier to only select brain MRI slices with/without tumors.

Towards DA, whereas realistic images give more insights on geometry/intensity transformations in classification, more realistic images do not always assure better DA, so we have to find suitable image resolutions and sequences; that is why we generate both high-resolution images and Concat images, yet they looked more unrealistic for the physician. For physician training, generating desired realistic tumors by adding conditioning requires exploring latent spaces of GANs extensively.

Overall, our novel GAN-based realistic brain MR image generation approach sheds light on diagnostic and prognostic medical applications; future studies on these applications are needed to confirm our encouraging results.

Chapter 5

GAN-based Medical Image Augmentation for 2D Classification

5.1 Prologue to Second Project

5.1.1 Project Publications

- **Infinite Brain MR Images: PGGAN-based Data Augmentation for Tumor Detection.** C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, E. Pasero (eds.) *Neural Approaches to Dynamics of Signal Exchanges*, Springer, pp. 291–303, September 2019.
- **Combining Noise-to-Image and Image-to-Image GANs: Brain MR Image Augmentation for Tumor Detection.** C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, *IEEE Access*, pp. 156966–156977, October 2019.

5.1.2 Context

At the time we wrote the former paper, high-resolution (e.g., 256×256) medical image generation using noise-to-image GANs had been challenging [72] while most CNN architectures adopt around 256×256 input sizes (e.g., InceptionResNetV2 [94]:

299×299 , ResNet-50 [95]: 224×224). Moreover, prior to the latter paper, analysis had been immature on GAN-generated additional training images for better CNN-based classification.

5.1.3 Contributions

This project’s core contribution is to firstly combine noise-to-image and image-to-image GANs for improved 2D classification. The former paper adopts a noise-to-image GAN called PGGANs to generate realistic/diverse original-sized 256×256 whole brain MR images with/without tumors separately; additionally, the latter paper exploits an image-to-image GAN called MUNIT to further refine the synthetic images’ texture and shape similarly to real ones. By so doing, our two-step GAN-based DA boosts sensitivity 93.7% to 97.5% in tumor/non-tumor classification. Moreover, we firstly analyze how medical GAN-based DA is associated with pre-training on ImageNet and discarding weird-looking synthetic images to humans to achieve high sensitivity. A physician classifies a few synthetic images as real in Visual Turing Test despite the high resolution.

5.1.4 Recent Developments

Since the former paper’s acceptance (the book chapter’s publication process took more than a year), to improve 2D classification, Konidaris *et al.* generated 192×160 brain MR images with Alzheimer’s disease using the noise-to-image GAN [73]. No more recent developments to report exist for the latter paper because it is very recent.

5.2 Motivation

CNNs are playing a key role in Medical Image Analysis, updating the state-of-the-art in many tasks [86, 96, 97] when large-scale annotated training data are available. However, preparing such massive medical data is demanding; thus, for better diagnosis, researchers generally adopt classic DA techniques, such as geometric/intensity transformations of original images [88, 89]. Those augmented images, however, intrinsically have a similar distribution to the original ones, resulting in limited performance improvement. In this sense, GAN-based DA can considerably increase the performance [7]; since the generated images are realistic but completely novel samples, they can relieve the sampling biases and fill the real image distribution uncovered by the original dataset [98].

The main problem in CAD lies in small/fragmented medical imaging datasets from multiple scanners; thus, researchers have improved classification by augmenting images with noise-to-image GANs [11] or image-to-image GANs [69]. However, no research has achieved further performance boost by combining noise-to-image and image-to-image GANs.

So, how can we maximize the DA effect under limited training images using the GAN combinations? To generate and refine brain MR images with/without tumors separately (Fig. 5-1), we propose a two-step GAN-based DA approach: (i) PGGANs [17], 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 (e.g., InceptionResNetV2 [94]: 299×299 , ResNet-50 [95]: 224×224); (ii) MUNIT [18] that combines GANs/VAEs [35] 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. Since training a single sophisticated GAN system is already difficult, instead of end-to-end training, we adopt a two-step approach for performance boost *via* an ensemble generation process from those state-of-the-art GANs' different algorithms.

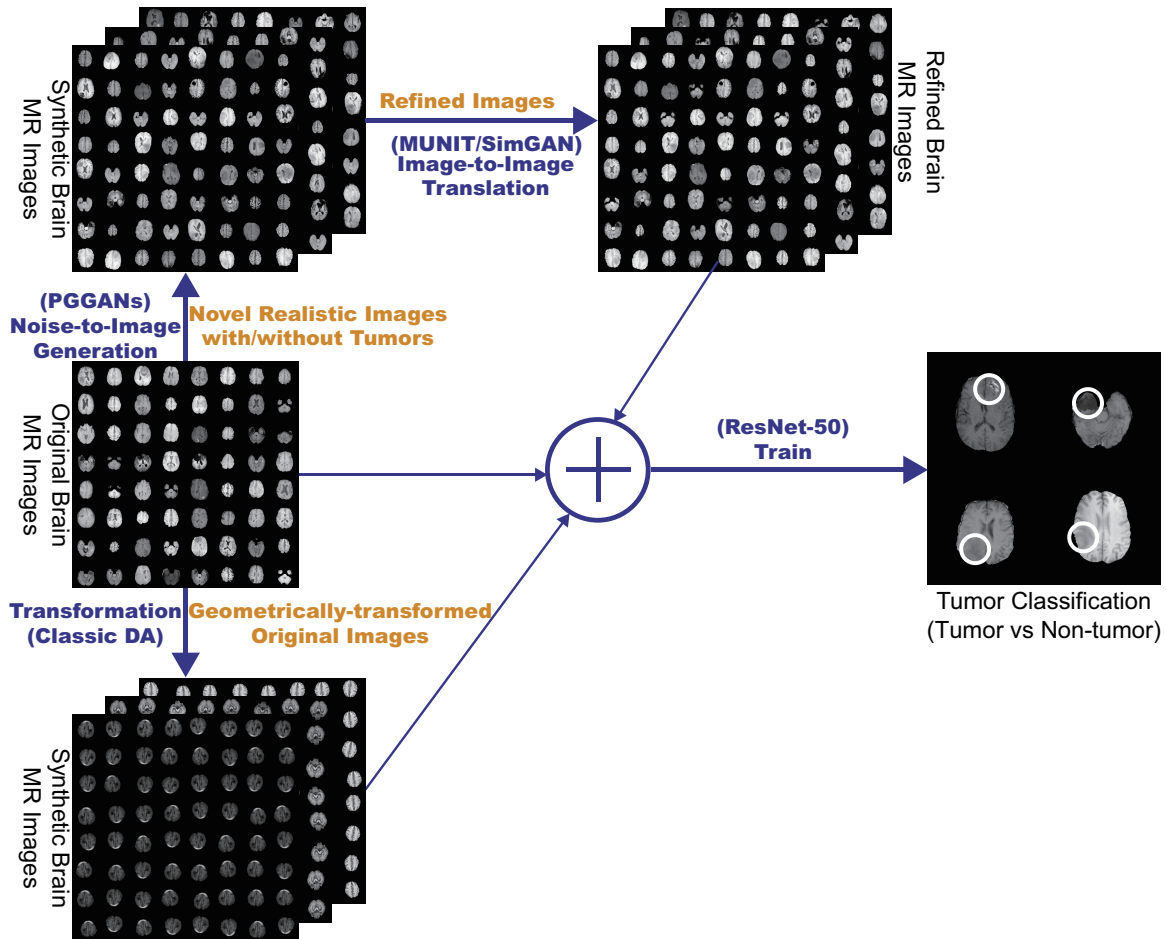


Figure 5-1: Combining noise-to-image and image-to-image GANs for better tumor classification: the PGGANs generates a number of realistic brain tumor/non-tumor MR images separately, the MUNIT/SimGAN refines them separately, and the binary classifier uses them as additional training data.

We thoroughly investigate CNN-based tumor classification results, also considering the influence of pre-training on ImageNet [34] and discarding weird-looking GAN-generated images. Moreover, we evaluate the synthetic images' appearance *via* Visual Turing Test [91] by an expert physician, and visualize the data distribution of real/synthetic images *via* t-Distributed Stochastic Neighbor Embedding (t-SNE) [99]. When combined with classic DA, our two-step GAN-based DA approach significantly outperforms the classic DA alone, boosting sensitivity 93.67% to 97.48%.

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?

Contributions. Our main contributions are as follows:

- **Whole Image Generation:** This research shows that PGGANs can generate realistic/diverse 256×256 whole medical images—not only small pathological sub-areas—and MUNIT can further refine their texture and shape similarly to real ones.
- **Two-step GAN-based DA:** This novel two-step approach, combining for the first time noise-to-image and image-to-image GANs, significantly boosts tumor 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.

5.3 Materials and Methods

5.3.1 BRATS 2016 Dataset

We use a dataset of 240×240 T1c brain axial MR images of 220 HGG cases from BRATS 2016 [92]. T1c is the most common sequence in tumor classification thanks to its high-contrast [100].

5.3.2 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, 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

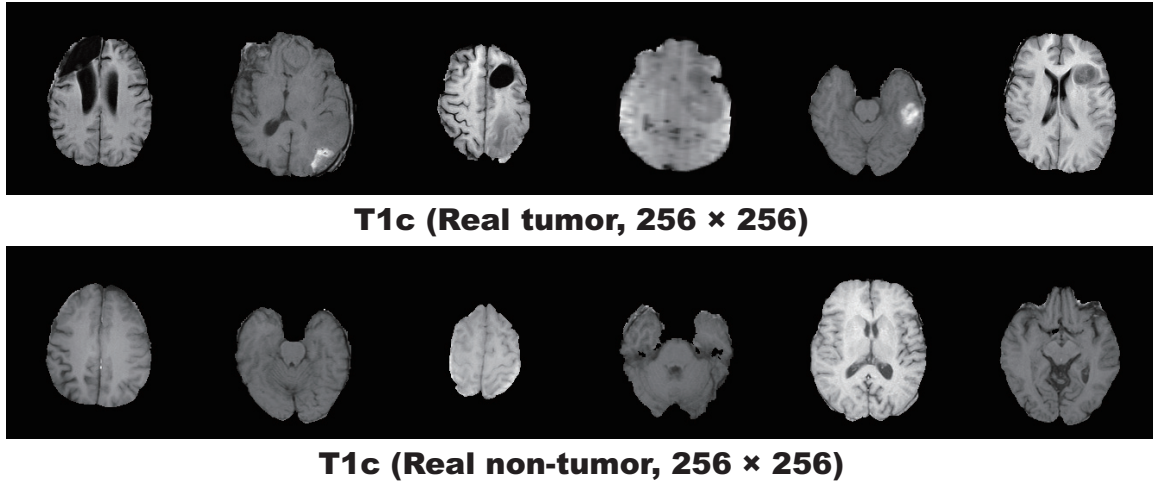


Figure 5-2: Example real 256×256 MR images used for PGGAN training.

(*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. 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. 5-2 shows example real MR images.

PGGANs [17] is a GAN training method that progressively grows a generator and discriminator: starting from low resolution, new layers model details as training progresses. This study adopts the PGGANs to synthesize realistic/diverse 256×256 brain MR images (Fig. 5-3); we train and generate tumor/non-tumor images separately.

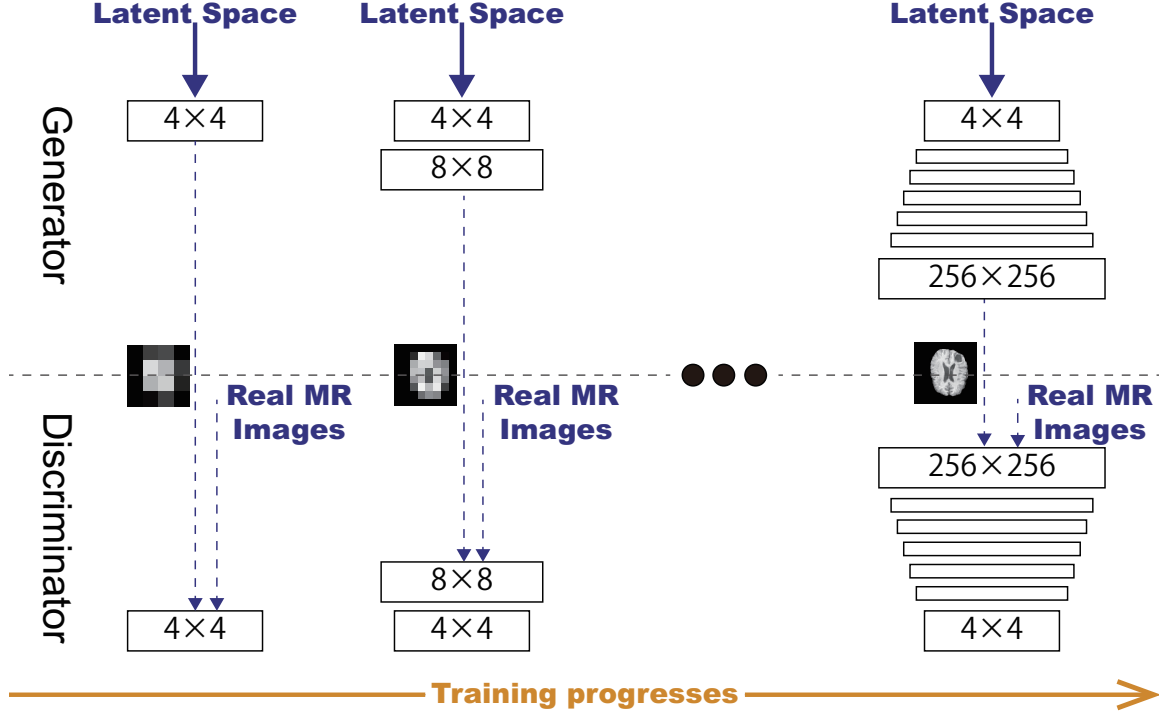


Figure 5-3: PGGAN architecture for 256×256 brain MR image generation. $N \times N$ refers to convolutional layers operating on $N \times N$ spatial resolution.

PGGAN Implementation Details The PGGAN architecture adopts the Wasserstein loss with Gradient Penalty (WGAN-GP) [39]:

$$\mathbb{E}_{\tilde{\mathbf{y}} \sim \mathbb{P}_g} [D(\tilde{\mathbf{y}})] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_r} [D(\mathbf{y})] + \lambda_{\text{gp}} \mathbb{E}_{\hat{\mathbf{y}} \sim \mathbb{P}_{\hat{\mathbf{y}}}} [(\|\nabla_{\hat{\mathbf{y}}} D(\hat{\mathbf{y}})\|_2 - 1)^2], \quad (5.1)$$

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

We train the model (Table 5.1) for 100 epochs with a batch size of 16 and 1.0×10^{-3} learning rate for the Adam optimizer (the exponential decay rates $\beta_1 = 0, \beta_2 = 0.99$) [101]. All experiments use $\lambda_{\text{gp}} = 10$ with 1 critic iteration per generator iteration. During training, we apply random cropping in 0-15 pixels as DA.

Table 5.1: PGGAN architecture details for the generator/discriminator. Pixelwise feature vector normalization [102] is applied in the generator after each convolutional layer except for the final output layer as in the original paper [17]. LReLU denotes Leaky ReLU with leakiness 0.2.

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

5.3.3 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 [18] or SimGAN [8]. SimGAN remarkably improved eye gaze estimation results after refining non-GAN-based synthetic images from the UnityEyes simulator *via* image-to-image translation; thus, we also expect such performance improvement after refining synthetic images from a noise-to-image GAN (i.e., PGGANs) *via* an image-to-image GAN (i.e., MUNIT/SimGAN) with considerably different GAN algorithms.

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

with the highest accuracy on tumor classification validation, when pre-trained and combined with classic DA, among 20,000/50,000/100,000 steps, respectively.

MUNIT [18] is an image-to-image GAN based on both auto-encoding/translation; it extends UNIT [103] to increase the generated images’ realism/diversity *via* a stochastic model representing continuous output distributions.

MUNIT Implementation Details The MUNIT architecture adopts the following loss:

$$\begin{aligned} \min_{E_1, E_2, G_1, G_2} \max_{D_1, D_2} & \mathcal{L}_{\text{VAE}_1} + \mathcal{L}_{\text{GAN}_1} + \mathcal{L}_{\text{CC}_1} + \mathcal{L}_{\text{VGG}_1} \\ & + \mathcal{L}_{\text{VAE}_2} + \mathcal{L}_{\text{GAN}_2} + \mathcal{L}_{\text{CC}_2} + \mathcal{L}_{\text{VGG}_2}, \end{aligned} \quad (5.2)$$

where $\mathcal{L}(\cdot)$ denotes the loss function. Using the multiple encoders E_1/E_2 , generators G_1/G_2 , discriminators D_1/D_2 , cycle-consistencies CC_1/CC_2 , and domain-invariant perceptions $\text{VGG}_1/\text{VGG}_2$ [104], this framework jointly solves learning problems of the $\text{VAE}_1/\text{VAE}_2$ and $\text{GAN}_1/\text{GAN}_2$ for the image reconstruction streams, image translation streams, cycle-consistency reconstruction streams, and domain-invariant perception streams. Since we do not need the style loss for our experiments, instead of the MUNIT loss, we use the UNIT loss with the perceptual loss for the MUNIT architecture (as in the UNIT authors’ GitHub repository). The MUNIT architecture adopts the following loss:

$$\begin{aligned} \min_{\epsilon, \tilde{\epsilon}, \delta} \max_{\tilde{\delta}, D, \tilde{D}} & \mathcal{L}_{\text{GAN}}^{\text{x}} + \mathcal{L}_{\text{GAN}}^{\tilde{\text{x}}} + \lambda_x (\mathcal{L}_{\text{recon}}^{\text{x}} + \mathcal{L}_{\text{recon}}^{\tilde{\text{x}}}) \\ & + \lambda_c (\mathcal{L}_{\text{recon}}^{\text{c}} + \mathcal{L}_{\text{recon}}^{\tilde{\text{c}}}), \end{aligned} \quad (5.3)$$

We train the model (Table 5.2) for 100,000 steps with a batch size of 1 and 1.0×10^{-4} learning rate for the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) [101]. The learning rate is reduced by half every 20,000 steps. We use the following MUNIT weights: the adversarial loss weight = 1; the image reconstruction loss weight = 10; the Kullback-Leibler (KL) divergence loss weight for reconstruction = 0.01; the cycle

Table 5.2: MUNIT architecture details for the generator/discriminator. We input color images (i.e., 3 channels) to use ImageNet initialization. Instance normalization [106]/adaptive instance normalization [107] are applied in the content encoder/decoder after each convolutional layer respectively except for the final decoder output layer as in the original paper [18]. LReLU denotes Leaky ReLU with leakiness 0.2.

Generator			Discriminator	Activation	Output Shape
Content Encoder			Input image	-	$3 \times 224 \times 224$
Input image	-	$3 \times 224 \times 224$	Conv 4×4	LReLU	$64 \times 112 \times 112$
Conv 7×7	ReLU	$64 \times 224 \times 224$	Conv 4×4	LReLU	$128 \times 56 \times 56$
Conv 4×4	ReLU	$128 \times 112 \times 112$	Conv 4×4	LReLU	$256 \times 28 \times 28$
Conv 4×4	ReLU	$256 \times 56 \times 56$	Conv 4×4	LReLU	$512 \times 14 \times 14$
ResBlock $\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 4$	ReLU	$256 \times 56 \times 56$	Conv 4×4	-	$1 \times 14 \times 14$
	-	$256 \times 56 \times 56$	AveragePool	-	$3 \times 112 \times 112$
Decoder			Conv 4×4	LReLU	$64 \times 56 \times 56$
ResBlock $\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 4$	ReLU	$256 \times 56 \times 56$	Conv 4×4	LReLU	$128 \times 28 \times 28$
	-	$256 \times 56 \times 56$	Conv 4×4	LReLU	$256 \times 14 \times 14$
Upsample	-	$256 \times 112 \times 112$	Conv 4×4	LReLU	$512 \times 7 \times 7$
Conv 5×5	ReLU	$128 \times 112 \times 112$	Conv 4×4	-	$1 \times 7 \times 7$
Upsample	-	$128 \times 224 \times 224$	AveragePool	-	$3 \times 56 \times 56$
Conv 5×5	ReLU	$64 \times 224 \times 224$	Conv 4×4	LReLU	$64 \times 28 \times 28$
Conv 7×7	Tanh	$3 \times 224 \times 224$	Conv 4×4	LReLU	$128 \times 14 \times 14$
			Conv 4×4	LReLU	$256 \times 7 \times 7$
			Conv 4×4	LReLU	$512 \times 3 \times 3$
			Conv 4×4	-	$1 \times 3 \times 3$
			AveragePool	-	$3 \times 28 \times 28$

consistency loss weight = 10; the KL divergence loss weight for cycle consistency = 0.01; the domain-invariant perceptual loss weight = 1; the Least Squares GAN objective function for the discriminators [105]. 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 The SimGAN architecture (i.e., a refiner) uses the following loss:

$$\sum_i \mathcal{L}_{\text{real}}(\boldsymbol{\theta}; \mathbf{x}_i, \mathcal{Y}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\boldsymbol{\theta}; \mathbf{x}_i), \quad (5.4)$$

where $\mathcal{L}(\cdot)$ denotes the loss function, $\boldsymbol{\theta}$ is the function parameters, \mathbf{x}_i is the i^{th} PGGAN-generated training image, and \mathcal{Y} is the set of the real images \mathbf{y}_j . The

Table 5.3: SimGAN architecture details for the refiner/discriminator. Batch normalization is applied both in the refiner/discriminator after each convolutional layer except for the final output layers respectively as in the original paper [8].

Refiner	Activation	Output Shape	Discriminator	Activation	Output Shape
Input image	–	$1 \times 224 \times 224$	Input image	–	$1 \times 224 \times 224$
Conv 9×9	ReLU	$64 \times 224 \times 224$	Conv 9×9	ReLU	$96 \times 72 \times 72$
ResBlock $\begin{bmatrix} 3 \times 3 \\ 3 \times 3 \end{bmatrix} \times 12$	ReLU	$64 \times 224 \times 224$	Conv 5×5	ReLU	$64 \times 68 \times 68$
	–	$64 \times 224 \times 224$	Maxpool	–	$64 \times 34 \times 34$
Conv 1×1	Tanh	$1 \times 224 \times 224$	Conv 5×5	ReLU	$64 \times 15 \times 15$
			Conv 3×3	ReLU	$32 \times 13 \times 13$
			Maxpool	–	$32 \times 7 \times 7$
			Conv 1×1	ReLU	$32 \times 7 \times 7$
			Conv 1×1	ReLU	$2 \times 7 \times 7$

first part $\mathcal{L}_{\text{real}}$ adds realism to the synthetic images using a discriminator, while the second part \mathcal{L}_{reg} preserves the tumor/non-tumor features.

We train the model (Table 5.3) for 20,000 steps with a batch size of 10 and 1.0×10^{-4} learning rate for the Stochastic Gradient Descent (SGD) optimizer [108] without momentum. The learning rate is reduced by half at 15,000 steps. We train the refiner first with just the self-regularization loss with $\lambda_{\text{reg}} = 5 \times 10^{-5}$ for 500 steps; then, for each update of the discriminator, we update the refiner 5 times. During training, we apply horizontal flipping as DA.

5.3.4 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 [95] is a 50-layer residual learning-based CNN. We adopt it to conduct tumor/non-tumor binary classification on MR images due to its outstanding performance in image classification tasks [109], including binary classification [24]. Chang *et al.*[110] also used a similar 34-layer residual convolutional network for the binary classification of brain tumors (i.e., determining the Isocitrate Dehydrogenase status in LGG/HGG).

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 [34] 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;
8. + 200k classic DA & 200k MUNIT-refined DA;
9. + 200k SimGAN-refined DA;
10. + 200k classic DA & 200k SimGAN-refined DA.

Due to the risk of overlooking the tumor diagnosis, higher sensitivity matters much more than higher specificity [111]; thus, we aim to achieve higher sensitivity, using the additional synthetic training images. We perform McNemar’s test on paired tumor classification results [112] to confirm our two-step GAN-based DA’s statistically-significant sensitivity improvement; since this statistical analysis involves multiple comparison tests, we adjust their p -values using the Holm-Bonferroni method [113].

Whereas medical imaging researchers widely use the ImageNet initialization despite different textures of natural/medical images, recent study found that such ImageNet-trained CNNs are biased towards recognizing texture rather than shape [114]; thus, we aim to investigate how the medical GAN-based DA affects classification performance with/without the pre-training. As the classic DA, we adopt a random combination of horizontal/vertical flipping, rotation up to 10 degrees, width/height shift up to 8%, shearing up to 8%, zooming up to 8%, and constant filling of points outside the input boundaries (Fig. 5-4). For the PGGAN-based DA and its refinement, we only use success cases after discarding weird-looking synthetic images (Fig. 5-5); DenseNet-169 [115] extracts image features and k-means++ [116] clusters the features into 200

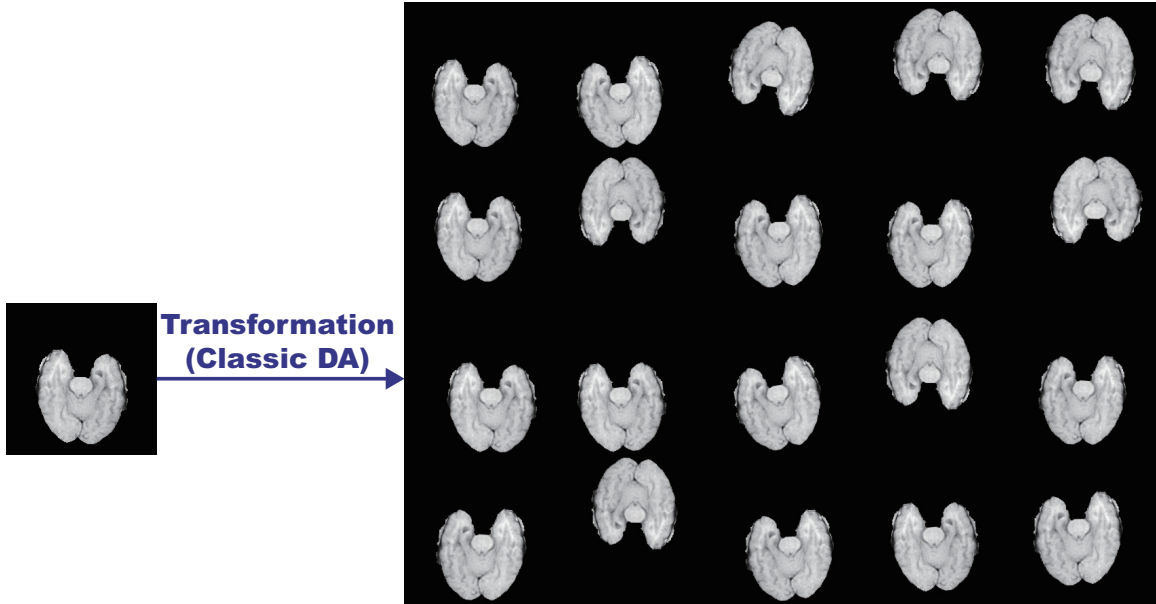


Figure 5-4: Example real 256×256 MR image and its geometrically-transformed images.

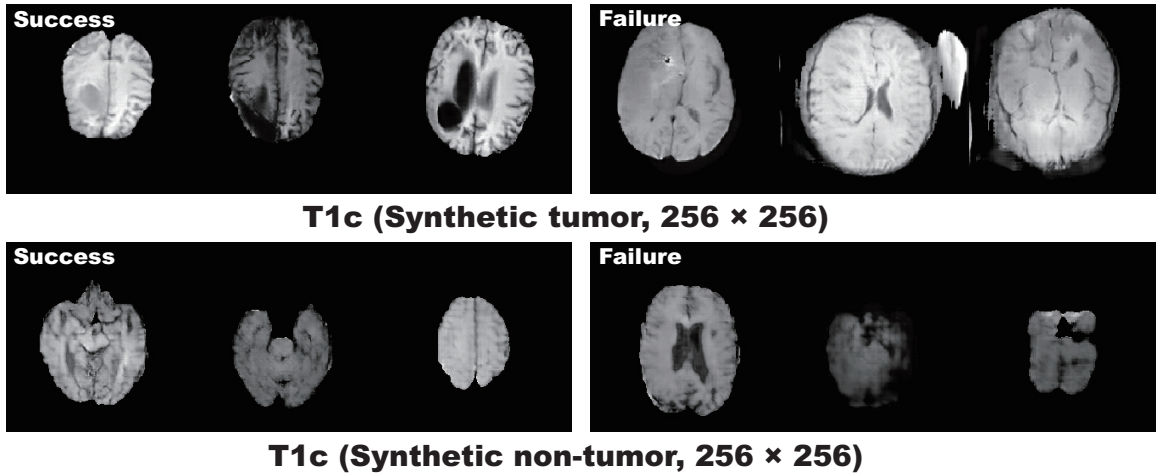


Figure 5-5: Example PGGAN-generated MR images: (a) Success cases; (b) Failure cases.

groups, and then we manually discard each cluster containing similar weird-looking images. To verify its effect, we also conduct a PGGAN-based DA experiment without the discarding step. Additionally, to confirm the effect of changing training data set sizes, we compare classification results with pre-training on 8,429/4,183/1,646/834 real images *vs* real images + 200k classic DA *vs* real images + 200k classic DA & 200k PGGAN-based DA (i.e., $4 \times 3 = 12$ setups).

ResNet-50 Implementation Details The ResNet-50 architecture adopts the binary cross-entropy loss for binary classification both with/without ImageNet pre-

Table 5.4: ResNet-50 architecture details without/with pre-training. We input grayscale images (i.e., 1 channel) for experiments without pre-training, whereas we input color images (i.e., 3 channels) for experiments with pre-training to use ImageNet initialization. Batch normalization is applied after each convolutional layer as in the original paper [95].

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

training. As shown in Table 5.4, for robust training, before the final sigmoid layer, we introduce a 0.5 dropout [57], linear dense, and batch normalization [117] 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 [108] 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.

5.3.5 Clinical Validation *via* Visual Turing Test

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

a random selection of:

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

Then, the physician is asked to classify them as both (*i*) real/synthetic and (*ii*) tumor/non-tumor, without previously knowing which is real/synthetic and tumor/non-tumor.

5.3.6 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-SNE [99] 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. We input the images after normalizing pixel values to $[0, 1]$. For point locations of the real images, we compress all the images simultaneously and plot each setup (i.e., the geometrically-transformed or each GAN-based images against the real ones) separately; we maintain their locations by projecting all the data onto the same subspace.

5.4 Results

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

5.4.1 MR Images Generated by PGGANs

Fig. 5-5 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).

5.4.2 MR Images Refined by MUNIT/SimGAN

MUNIT and SimGAN differently refine PGGAN-generated images—they render the texture and contours while maintaining the overall shape (Fig. 5-6). Non-tumor images change more remarkably than tumor images for both MUNIT and SimGAN; it probably derives from unsupervised image translation’s loss for consistency to avoid image collapse, resulting in conservative change for more complicated images.

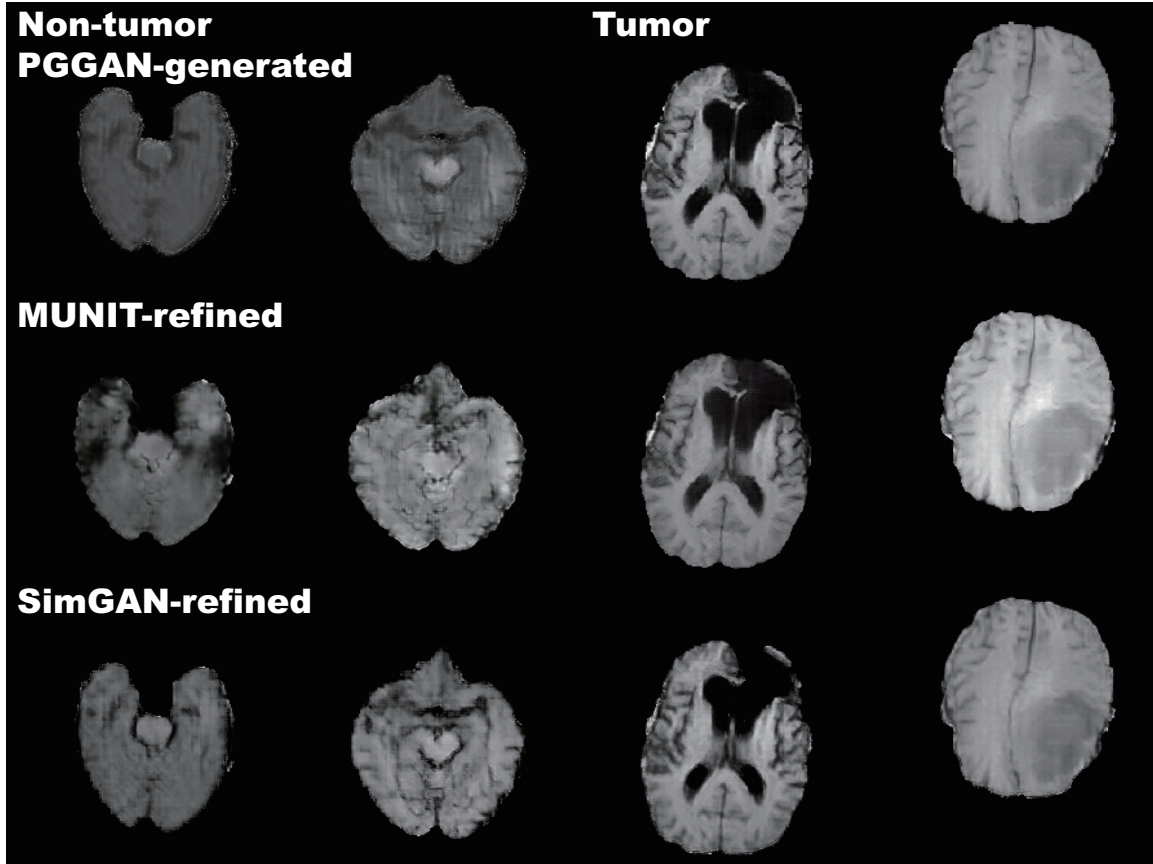


Figure 5-6: Example PGGAN-generated 256×256 MR images and their refined versions by MUNIT/SimGAN.

5.4.3 Tumor Classification Results

Table 5.5 shows the brain tumor classification results with/without DA while Table 5.6 indicates their pairwise comparison (p -values between our two-step GAN-based DA setups and the other DA setups) using McNemar’s test. ImageNet pre-training generally outperforms training from scratch despite different image domains (i.e., natural images to medical images). As expected, classic DA remarkably improves classification, while no clear difference exists between the 200,000/400,000 classic DA under sufficient geometrically-transformed training images. When pre-trained, each GAN-based DA (i.e., PGGANs/MUNIT/SimGAN) alone helps classification due to the robustness from GAN-generated images; but, without pre-training, it harms classification due to the biased initialization from the GAN-overwhelming data distribution. Similarly, without pre-training, PGGAN-based DA without clustering/discarding causes poor

Table 5.5: ResNet-50 tumor results of 20 DA setups, with (without) ImageNet pre-training. Sensitivity and specificity consider the slight tumor/non-tumor class imbalance (about 6 : 5) in the test set. Boldface indicates the best performance.

DA Setups	Accuracy (%)	Sensitivity (%)	Specificity (%)
(1) 8,429 real images	93.1 (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)

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 remarkably outperforms the GAN-based DA or classic DA alone in terms of sensitivity since they are mutually-complementary: the former learns the non-linear manifold of the real images to generate novel local tumor features (since we train tumor/non-tumor images separately) strongly associated with sensitivity; the latter learns the geometrically-transformed manifold of the real images to cover global features and provide the robustness on training for most cases. We confirm that test samples, originally-misclassified but correctly classified after DA, are obviously different for the GAN-based DA and classic DA; here, both image-to-image GAN-based DA, especially MUNIT, produce remarkably higher sensitivity than the PGGAN-based DA after refinement. Specificity is higher than sensitivity for every DA setup with pre-training, probably due to the training data imbalance; but interestingly, without pre-training, sensitivity is higher than specificity for both image-to-image GAN-based DA since our tumor classification-oriented two-step GAN-based DA can fill the real tumor image distribution uncovered by the original dataset under no ImageNet initialization. Accordingly, when combined with the classic DA, the MUNIT-based DA based on both GANs/VAEs achieves the highest sensitivity 97.48% against the best performing clas-

Table 5.6: McNemar’s test p -values for the pairwise comparison of the ResNet-50 tumor classification results by accuracy, sensitivity, and specificity. We compare our two-step GAN-based DA setups and all the other DA setups. All numbers within parentheses refer to DA setups on Table 5.5 and PT denotes pre-training. Boldface indicates statistical significance (threshold p -value < 0.05).

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

sic DA’s 93.67%, allowing to significantly alleviate the risk of overlooking the tumor diagnosis; in terms of sensitivity, it outperforms all the other DA setups, including two-step DA setups, with statistical significance.

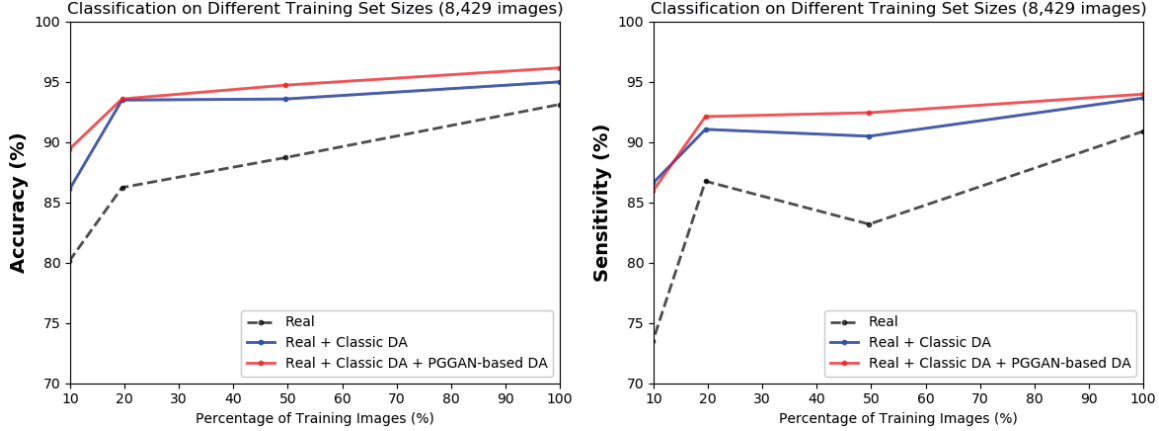


Figure 5-7: ResNet-50 tumor classification results under different training dataset sizes of 12 DA setups, with ImageNet pre-training: (a) 8,429/4,183/1,646/834 real images *vs* (b) real images + 200k Classic DA *vs* (c) real images + 200k Classic DA & 200k PGGAN-based DA.

Figure 5-7 shows that the PGGAN-based DA, even without further refinement, could moderately increase both accuracy/sensitivity on top of the classic DA in tumor classification; it achieves considerably high sensitivity with only 20%/50% of the real training images. However, it should be noted that the MUNIT-based DA could outperform the PGGAN-based DA in return for more computational power.

5.4.4 Visual Turing Test Results

Table 5.7 indicates the confusion matrix for the Visual Turing Test. The expert physician classifies a few PGGAN-generated images as real, thanks to their realism, despite high resolution (i.e., 224×224 pixels); meanwhile, the expert classifies less GAN-refined images as real due to slight artifacts induced during refinement. The synthetic images successfully capture tumor/non-tumor features; unlike the non-tumor images, the expert recognizes a considerable number of the mild/modest tumor images as non-tumor for both real/synthetic cases. It derives from clinical tumor diagnosis relying on a full 3D volume, instead of a single 2D slice.

Table 5.7: Visual Turing Test results by an expert physician for classifying Real (R) *vs* Synthetic (S) images and Tumor (T) *vs* Non-tumor (N) images. Accuracy denotes the physician’s successful classification ratio between the real/synthetic images and between the tumor/non-tumor images, respectively. It should be noted that proximity to 50% of accuracy indicates superior performance (chance = 50%).

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

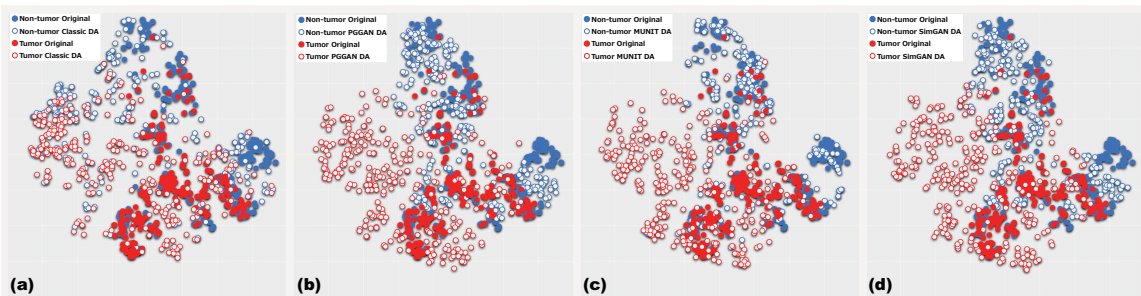


Figure 5-8: T-SNE plots with 300 tumor/non-tumor images per each category: Real images *vs* (a) Geometrically-transformed images; (b) PGGAN-generated images; (c) MUNIT-refined images; (d) SimGAN-refined images.

5.4.5 T-SNE Results

As Fig. 5-8 represents, the real tumor/non-tumor image distributions largely overlap while the non-tumor images distribute wider. The geometrically-transformed tumor/non-tumor image distributions also often overlap, and both images distribute wider than the real ones. All GAN-based synthetic images by PGGANs, MUNIT, and SimGAN distribute widely, while their tumor/non-tumor images overlap much less than the geometrically-transformed ones (i.e., a high discrimination ability associated with sensitivity improvement); the MUNIT-refined images show better tumor/non-tumor discrimination and a more similar distribution to the real ones than the

PGGAN-based and SimGAN-based images. This trend derives from the MUNIT’s loss function adopting both GANs/VAEs that further fits the PGGAN-generated images into the real image distribution by refining their texture and shape; contrarily, this refinement could also induce slight human-recognizable but DA-irrelevant artifacts. Overall, the GAN-based images, especially the MUNIT-refined images, fill the distribution uncovered by the real or geometrically-transformed ones with less tumor/non-tumor overlap; this demonstrates the superiority of combining classic DA and GAN-based DA.

5.5 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; thanks to an ensemble generation process from those GANs’ different algorithms, the texture/shape-refined images can replace missing data points of the training set with less tumor/non-tumor overlap, and thus handle the data imbalance by regularizing the model (i.e., improved generalization). Notably, MUNIT remarkably outperforms SimGAN in terms of sensitivity, probably due to the effect of combining both GANs/VAEs.

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

them boosts DA performance.

Overall, by minimizing the number of annotated images required for medical imaging tasks, the two-step GAN-based DA can shed light not only on classification, but also on object detection [20] and segmentation [76]. Moreover, other potential medical applications exist: (*i*) A data anonymization tool to share patients' data outside their institution for training without losing classification performance [76]; (*ii*) A physician training tool to show random pathological images for medical students/radiology trainees despite infrastructural/legal constraints [13]. As future work, we plan to define a new end-to-end GAN loss function that explicitly optimizes the classification results, instead of optimizing visual realism while maintaining diversity by combining the state-of-the-art noise-to-image and image-to-image GANs; towards this, we might extend a preliminary work on a three-player GAN for classification [118] to generate only hard-to-classify samples to improve classification; we could also (*i*) explicitly model deformation fields/intensity transformations and (*ii*) leverage unlabeled data during the generative process [26] to effectively fill the real image distribution.

Chapter 6

GAN-based Medical Image Augmentation for 2D Detection

6.1 Prologue to Third Project

6.1.1 Project Publication

- **Learning More with Less: Conditional PGGAN-based Data Augmentation for Brain Metastases Detection Using Highly-Rough Annotation on MR Images.** C. Han, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, L. Rundo, H. Nakayama, S. Satoh, In ACM International Conference on Information and Knowledge Management (CIKM), Beijing, China, pp. 119–127, November 2019.

6.1.2 Context

Further DA applications require pathology localization for detection and advanced physician training needs atypical image generation, respectively. To meet both clinical demands, developing pathology-aware GANs (i.e., GANs conditioned on pathology position and appearance) is the best solution—the pathology-aware GANs are promising in terms of extrapolation because common and/or desired medical priors can play a key role in the conditioning [19]. However, prior to this work, researchers

had focused only on improving segmentation, instead of bounding box-based detection, while the detection requires much less physicians' annotation effort [76, 77]. Moreover, they had relied on image-to-image GANs, instead of noise-to-image GANs, which sacrifices image diversity due to an input benign image.

6.1.3 Contributions

This project's fundamental contribution is to propose a novel pathology-aware noise-to-image GAN called CPGGANs for improved 2D bounding box-based detection; it incorporates highly-rough bounding box conditions incrementally into the noise-to-image GAN (i.e., PGGANs) to place realistic/diverse brain metastases at desired positions/sizes on 256×256 MR images. By so doing, our CPGGAN-based DA boosts sensitivity 83% to 91% with Intersection over Union (IoU) threshold 0.25 in tumor detection with clinically acceptable additional False Positives (FPs). Moreover, we find that GAN training on additional normal images could increase synthetic images' realism, including pathology, but decrease DA performance.

6.1.4 Recent Developments

Almost simultaneously, Kanayama *et al.* also tackle bounding box-based pathology detection using the image-to-image GAN, instead of the noise-to-image GAN [74]; 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 ROI at desired position.

6.2 Motivation

Accurate CAD with high sensitivity can alleviate the risk of overlooking the diagnosis in a clinical environment. Specifically, CNNs have revolutionized medical imaging, such as diabetic eye disease diagnosis [119], mainly thanks to large-scale annotated training data. However, obtaining such annotated medical big data is demanding; thus, better diagnosis requires intensive DA techniques, such as geometric/intensity transformations of original images [88, 89]. Yet, those augmented images intrinsically have a similar distribution to the original ones, leading to limited performance improvement; in this context, GAN [7]-based DA can boost the performance by filling the real image distribution uncovered by the original dataset, since it generates realistic but completely new samples showing good generalization ability; GANs achieved outstanding performance in computer vision, including 21% performance improvement in eye-gaze estimation [8].

Also in medical imaging, where the primary problem lies in small and fragmented imaging datasets from various scanners [96], GAN-based DA performs effectively: researchers improved classification by augmentation with noise-to-image GANs [12] and segmentation with image-to-image GANs [76, 77]. Such applications include 256×256 brain MR image generation for tumor/non-tumor classification [15]. Nevertheless, unlike bounding box-based object detection, simple classification cannot locate disease areas and rigorous segmentation requires physicians' expensive annotation.

So, 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. 6-1); this is clinically valuable for better diagnosis, prognosis, and treatment, since brain metastases are the most common intra-cranial tumors, getting prevalent as oncological therapies improve cancer patients' survival [120]. Conventional GANs cannot generate realistic 256×256 whole brain MR images conditioned on tumor positions/sizes under limited training data/highly-rough annotation [15]; since noise-to-image GANs cannot

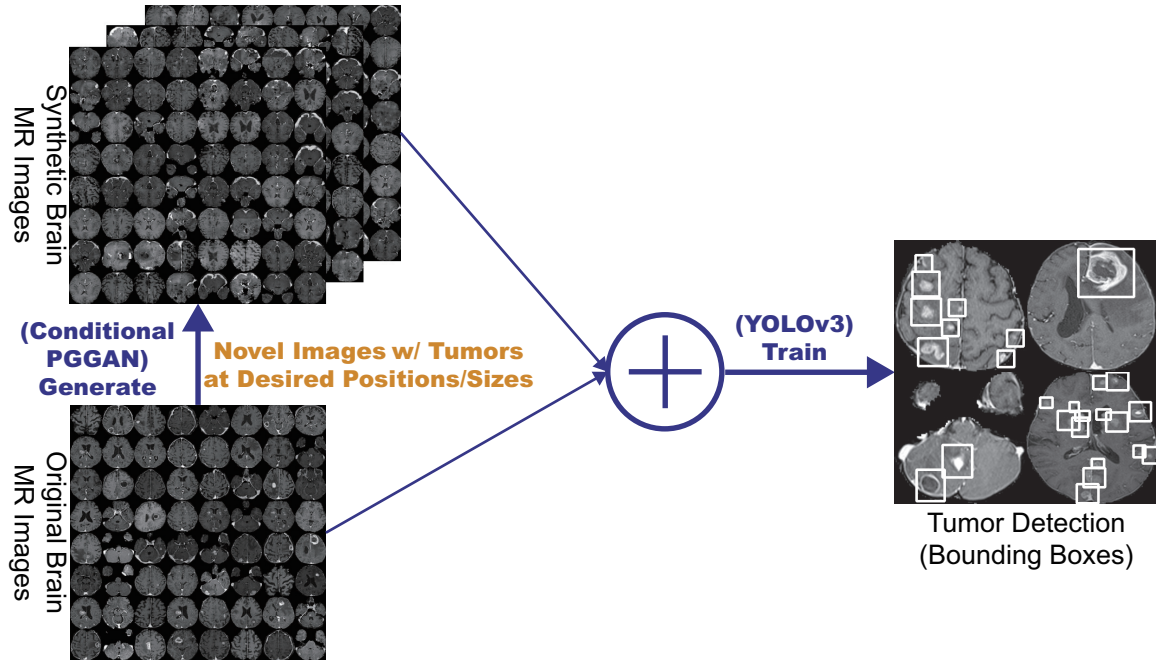


Figure 6-1: CPGGAN-based DA for better tumor detection: our CPGGANs generates a number of realistic/diverse brain MR images with tumors at desired positions/sizes based on bounding boxes, and the object detector uses them as additional training data.

directly be conditioned on an image describing desired objects, we have to use image-to-image GANs (e.g., input both conditioning image/random noise samples or the conditioning image alone with dropout noises [57] on a generator [42])—it results in unrealistic high-resolution MR images with odd artifacts due to the limited training data/rough annotation, tumor variations, and strong consistency in brain anatomy, unless we also input a benign image sacrificing image diversity.

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, as a conditional noise-to-image GAN not relying on an input benign image, we propose CPGGANs, incorporating highly-rough bounding box conditions incrementally into PGGANs [17] to naturally place tumors of random shape at desired positions/sizes on MR images. Moreover, we evaluate the generated images’ realism *via* Visual Turing Test [91] by three expert physicians, and visualize the data distribution *via* t-SNE algorithm [99]. Using the synthetic images, our novel CPGGAN-based DA boosts 10% sensitivity

in diagnosis with clinically acceptable additional FPs. Surprisingly, we confirm that further realistic tumor appearance, judged by the physicians, does not contribute to detection performance.

Research Questions. We mainly address two questions:

- **PGGAN Conditioning:** How can we modify PGGANs to naturally place objects of random shape, unlike rigorous segmentation, at desired positions/sizes based on highly-rough bounding box masks?
- **Medical DA:** How can we balance the number of real and additional synthetic training data to achieve the best detection performance?

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; the generated images can play a vital role in clinical oncology applications, such as DA, data anonymization, and physician training.
- **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.

6.3 Materials and Methods

6.3.1 Brain Metastases Dataset

As a new dataset for the first bounding box-based brain metastases detection, this project uses a dataset of T1c brain axial MR images, collected by the authors (National Center for Global Health and Medicine, Tokyo, Japan) and currently not publicly available for ethical restrictions; for robust clinical applications, it contains 180 brain metastatic cancer cases from multiple MRI scanners—those images differ in contrast, magnetic field strength (i.e., 1.5 T, 3.0 T), and matrix size (i.e., 190×224 , 216×256 , 256×256 , 460×460 pixels). 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.

6.3.2 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).

Our training set is relatively small/fragmented for CNN-based applications, considering that the same patient’s tumor slices could convey very similar information. 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. After skull-stripping on

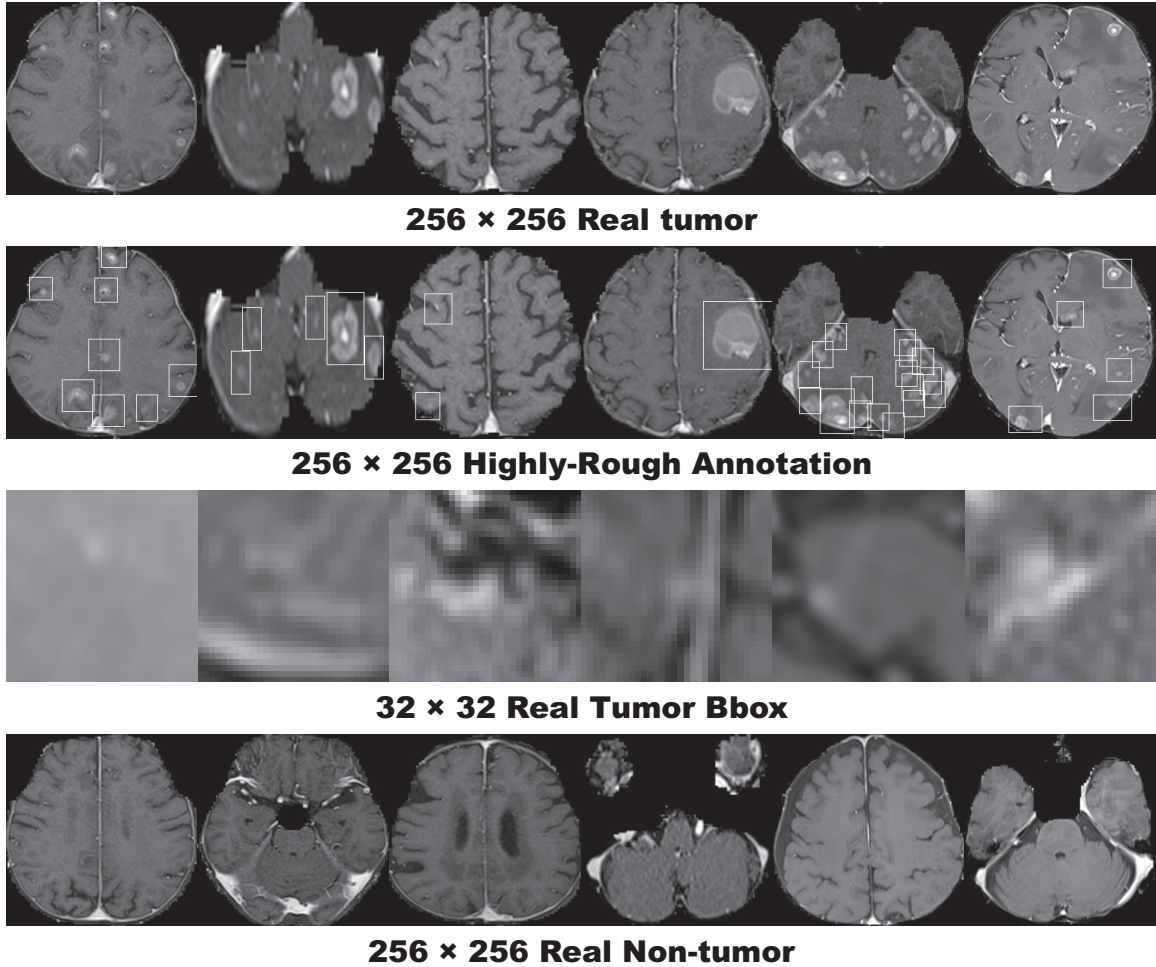


Figure 6-2: Example real 256×256 MR images with highly-rough annotation used for GAN training and resized 32×32 tumor bounding boxes.

all images with various resolution, remaining brain parts are cropped and resized to 256×256 pixels (i.e., a power of 2 for better GAN training). As Fig. 6-2 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 [17], unlike conditional image-to-image GANs requiring rigorous segmentation masks [121]. The original PGGANs exploits a progressively growing generator and discriminator: starting from low-resolution, newly-added layers model fine-grained details as training

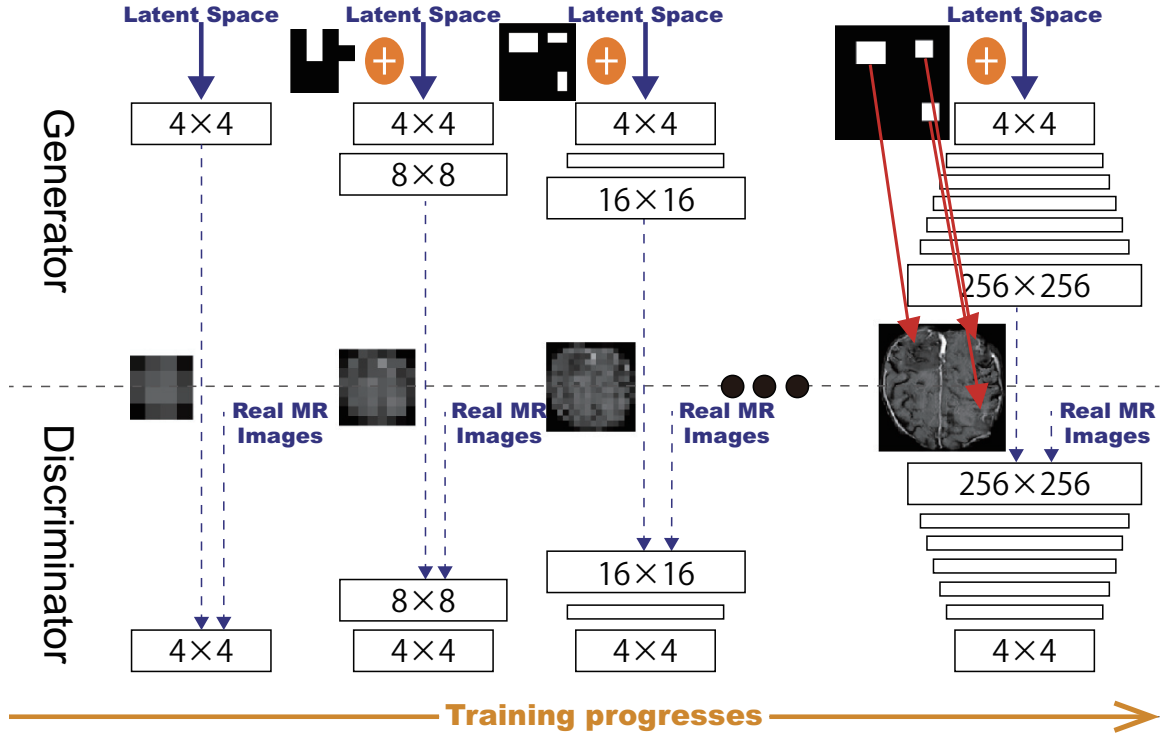


Figure 6-3: Proposed CPGGAN architecture for synthetic 256×256 brain MR image generation with tumors at desired positions/sizes based on bounding boxes.

progresses. As Fig. 6-3 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 without an input benign image under limited training data/highly-rough annotation. 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.

CPGGAN Implementation Details We use the CPGGAN architecture with the WGAN-GP loss [39]:

$$\mathbb{E}_{\tilde{y} \sim \mathbb{P}_g} [D(\tilde{y})] - \mathbb{E}_{y \sim \mathbb{P}_r} [D(y)] + \lambda \mathbb{E}_{\hat{y} \sim \mathbb{P}_{\hat{y}}} [(\|\nabla_{\hat{y}} D(\hat{y})\|_2 - 1)^2] \quad (6.1)$$

where the discriminator D belongs to 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 synthetic sample \tilde{y} generated from the conditioning image uniform noise samples in $[-1, 1]$. The last term is gradient penalty for the random sample $\hat{y} \sim \mathbb{P}_{\hat{y}}$.

Training lasts for 3,000,000 steps with a batch size of 4 and 2.0×10^{-4} learning rate for the Adam optimizer [101]. We flip the discriminator’s real/synthetic labels once in three times for robustness. 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 that generates brain MR images with tumors, concatenating a 256×256 conditioning image with noise samples for a generator input and concatenating the conditioning image with a real/synthetic image for a discriminator input, respectively. It uses a U-Net-like [88] generator with 4 convolutional/deconvolutional layers in encoders/decoders respectively with skip connections, along with a discriminator with 3 decoders. We apply batch normalization [117] to both convolution with LeakyReLU and deconvolution with ReLU. It follows the same implementation details as for the CPGGANs.

6.3.3 YOLOv3-based Brain Metastases Detection

You Only Look Once v3 (YOLOv3) [122] is a fast/accurate CNN-based object detector: unlike conventional classifier-based detectors, it divides the image into regions and predicts bounding boxes/probabilities for each region. We adopt YOLOv3 to detect brain metastases since its high efficiency can play a clinical role in real-

time tumor alert; moreover, it shows very comparable results with 608×608 network resolution against other state-of-the-art detectors, such as Faster RCNN [123].

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 a clinically acceptable number of FPs, adding the additional synthetic training images. Since our annotation is highly-rough, we calculate sensitivity/FPs per slice with both IoU threshold 0.5 and 0.25.

YOLOv3 Implementation Details We use the YOLOv3 architecture with Darknet-53 as a backbone classifier and sum squared error between the predictions/ground truth as a loss:

$$\begin{aligned}
& \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\
& + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \\
& + \sum_{i=0}^{S^2} \mathbb{1}_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \tag{6.2}
\end{aligned}$$

where x_i, y_i are the centroid location of an anchor box, w_i, h_i are the width/height of the anchor, C_i is the Objectness (i.e., confidence score of whether an object exists), and $p_i(c)$ is the classification loss. Let S^2 and B be the size of a feature map and the number of anchor boxes, respectively. $\mathbb{1}_i^{\text{obj}}$ is 1 when an object exists in cell i and otherwise 0.

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 to detect small tumors better. We recalculate the anchors at each DA setup. 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 while achieving high sensitivity.

6.3.4 Clinical Validation *via* Visual Turing Test

To quantitatively evaluate how realistic the CPGGAN-based synthetic images are, we supply, in random order, to three expert physicians a random selection of 50 real and 50 synthetic brain metastases images. They take four tests in ascending order: (i), (ii) test 1, 2: real *vs* CPGGAN-generated resized 32×32 tumor bounding boxes, trained without/with additional normal brain images; (iii), (iv) test 3, 4: real *vs* CPGGAN-generated 256×256 MR images, trained without/with additional normal brain images.

Then, the physicians constantly classify them as real/synthetic, if needed, zooming/rotating them, without previous training stages revealing which is real/synthetic.

6.3.5 Visualization *via* t-SNE

To visually analyze the distribution of real/synthetic images, we use t-SNE [99] on a random selection of:

- 500 real tumor images;
- 500 CPGGAN-generated tumor images;
- 500 CPGGAN-generated tumor images, trained with additional normal brain images.

We normalize the input images to $[0, 1]$.

t-SNE Implementation Details We use t-SNE with a perplexity of 100 for 1,000 iterations to get a 2D representation.

6.4 Results

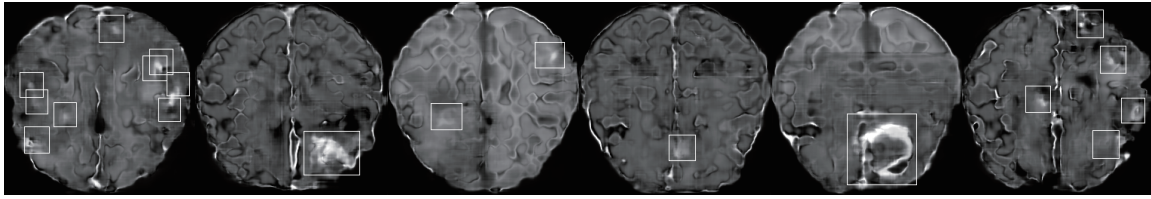
This section shows how CPGGANs and image-to-image GAN generate brain MR images. The results include instances of synthetic images and their influence on tumor detection, along with CPGGAN-generated images' evaluation *via* Visual Turing Test and t-SNE.

6.4.1 MR Images Generated by CPGGANs

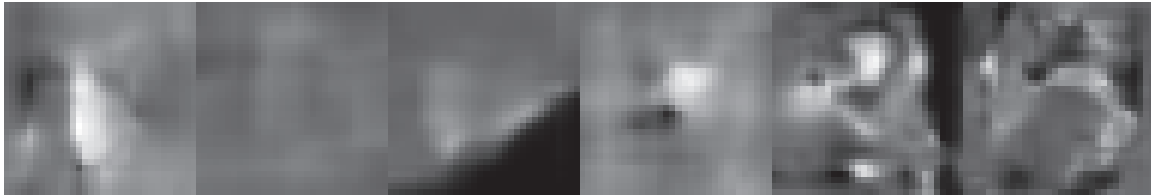
Fig. 6-4 illustrates example GAN-generated images. CPGGANs successfully captures the T1c-specific texture and tumor appearance at desired positions/sizes. Since we use highly-rough bounding boxes, the synthetic tumor shape largely varies within the boxes. When trained with additional normal brain images, it clearly maintains the realism of the original images with less odd artifacts, including tumor bounding boxes, which the additional images do not include. However, as expected, image-to-image GAN, without progressive growing, generates clearly unrealistic images without an input benign image due to the limited training data/rough annotation.

6.4.2 Brain Metastases Detection Results

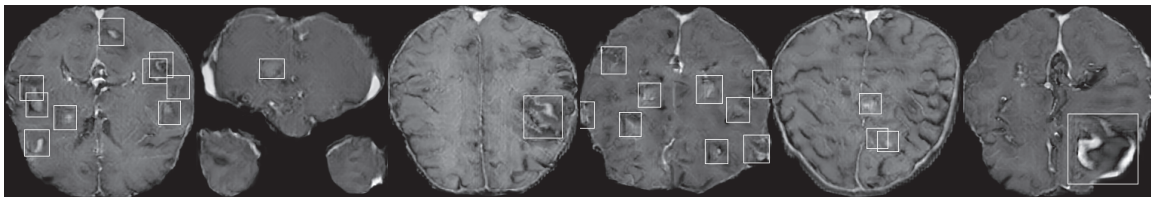
Table 6.1 shows the tumor detection results with/without GAN-based DA. As expected, the sensitivity remarkably increases with the additional synthetic training data while FPs per slice also increase. Adding more synthetic images generally leads to a higher amount of FPs, also detecting blood vessels that are small/hyper-intense on T1c MR images, very similarly to the enhanced tumor regions (i.e., the contrast agent perfuses throughout the blood vessels). However, 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, probably due to



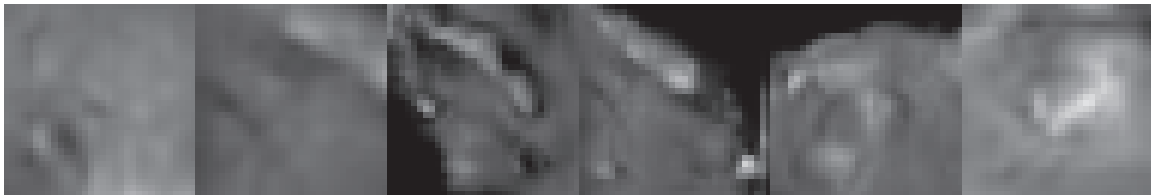
256 × 256 CPGGAN-generated Tumor w/o Normal



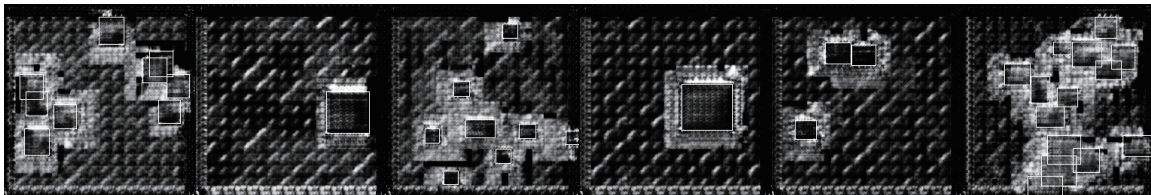
32 × 32 CPGGAN-generated Tumor Bbox w/o Normal



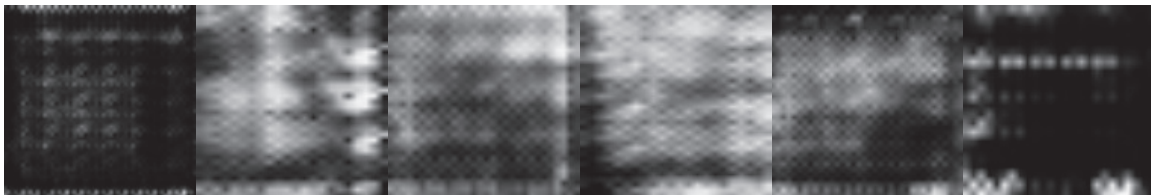
256 × 256 CPGGAN-generated Tumor w/ Normal



32 × 32 CPGGAN-generated Tumor Bbox w/ Normal



256 × 256 Image-to-Image GAN-generated Tumor w/o Normal



32 × 32 Image-to-Image GAN-generated Tumor Bbox w/o Normal

Figure 6-4: 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.

Table 6.1: 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

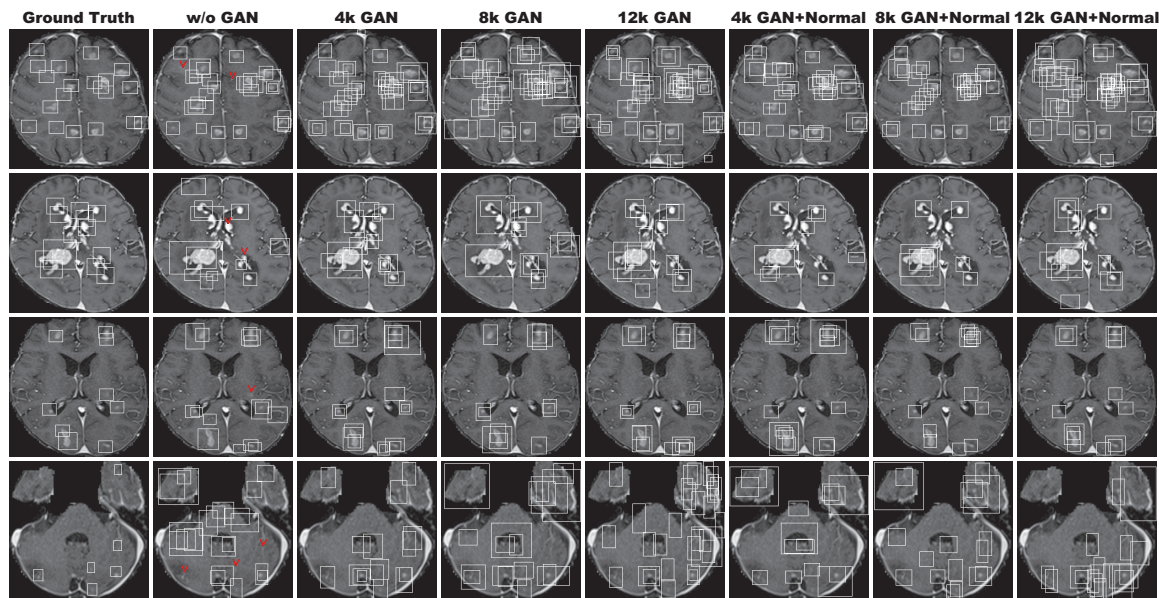


Figure 6-5: Example detection results of seven DA setups on four different images, compared against the ground truth: (a) ground truth; (b) without CPGGAN-based DA; (c), (d), (e) with 4k/8k/12k CPGGAN-based DA; (f), (g), (h) with 4k/8k/12k CPGGAN-based DA, trained with additional normal brain images. Red V symbols indicate the brain metastases undetected without CPGGAN-based DA, but detected with 4k CPGGAN-based DA.

the real/synthetic training image balance—the improved training robustness achieves sensitivity 0.91 with moderate IoU threshold 0.25 despite our highly-rough bounding box annotation.

Fig. 6-5 also visually indicates that it can alleviate the risk of overlooking the tumor diagnosis with clinically acceptable FPs; in the clinical routine, the bounding boxes, highly-overlapping around tumors, only require a physician’s single check by

Table 6.2: Visual Turing Test results by three physicians for classifying real *vs* CPGGAN-generated images: (a), (b) Test 1, 2: resized 32×32 tumor bounding boxes, trained without/with additional normal brain images; (c), (d) Test 3, 4: 256×256 MR images, trained without/with normal brain images. Accuracy denotes the physicians’ successful classification ratio between the real/synthetic images.

		Accuracy (%)	Real as Real (%)	Real as Synt (%)	Synt as Real (%)	Synt as Synt (%)
Test 1	Physician 1	88	80	20	4	96
	Physician 2	95	90	10	0	100
	Physician 3	97	98	2	4	96
Test 2	Physician 1	81	78	22	16	84
	Physician 2	83	86	14	20	80
	Physician 3	91	90	10	8	92
Test 3	Physician 1	97	94	6	0	100
	Physician 2	96	92	8	0	100
	Physician 3	100	100	0	0	100
Test 4	Physician 1	91	82	18	0	100
	Physician 2	96	96	4	4	96
	Physician 3	100	100	0	0	100

switching on/off transparent alpha-blended annotation on MR images. It should be noted that we cannot increase FPs to achieve such high sensitivity without CPGGAN-based DA. 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. Image-to-image GAN-based DA just moderately facilitates detection with less additional FPs, probably because the synthetic images have a distribution far from the real ones and thus their influence on detection is limited during testing.

6.4.3 Visual Turing Test Results

Table 6.2 shows the confusion matrix for the Visual Turing Test. The expert physicians easily recognize 256×256 synthetic images due to the lack of training data. However, when CPGGANs is trained with additional normal brain images, the experts classify a considerable number of synthetic tumor bounding boxes as real; it implies that the additional normal images remarkably facilitate the realism of both healthy and pathological brain parts while they do not include abnormality; thus, CPGGANs might perform as a tool to train medical students and radiology trainees when enough medical images are unavailable, such as abnormalities at rare posi-

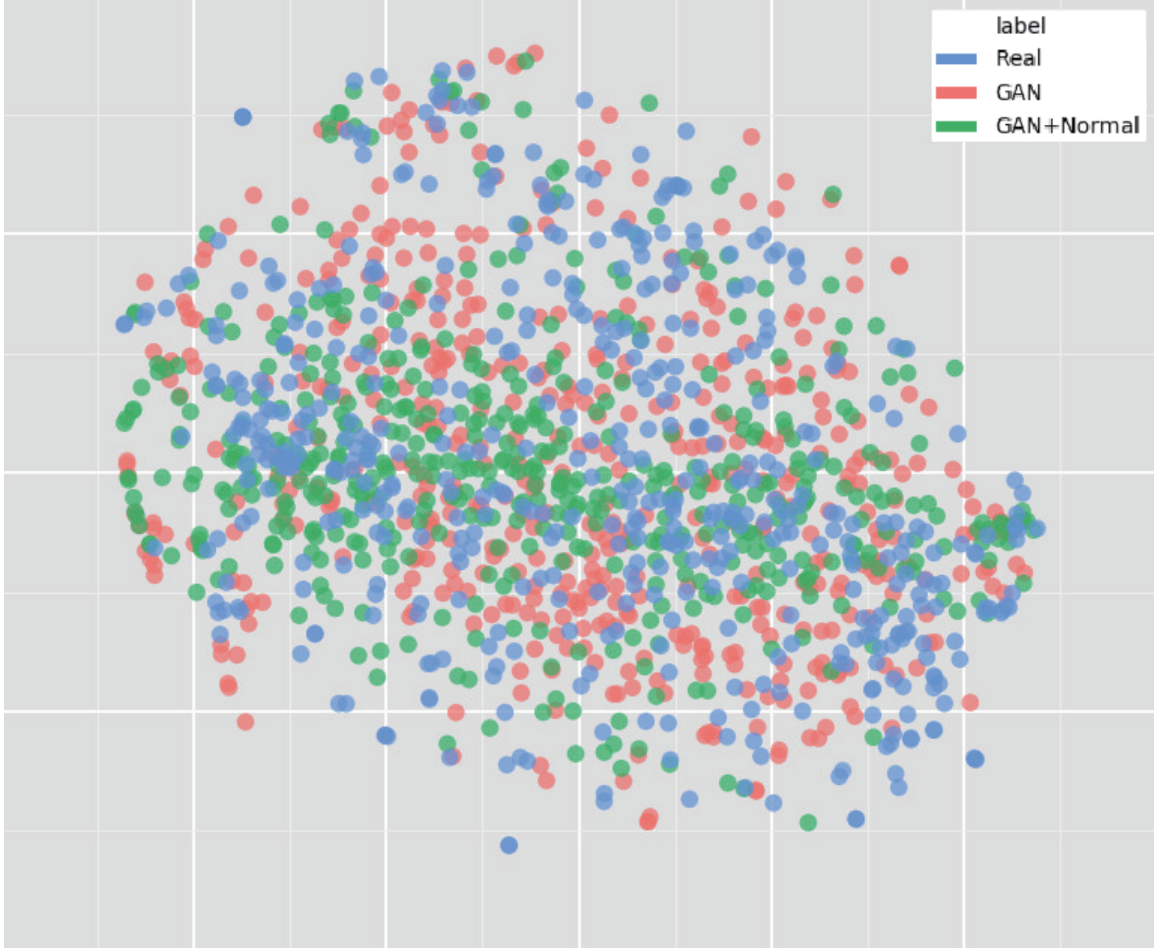


Figure 6-6: T-SNE results with 500 32×32 resized tumor bounding box images per each category: (a) Real tumor images; (b), (c) CPGGAN-generated tumor images, trained without/with additional normal brain images.

tions/sizes. Such GAN applications are clinically prospective [13], considering the expert physicians’ positive comments about the tumor realism.

6.4.4 T-SNE Results

As presented in Fig. 6-6, synthetic tumor bounding boxes have a moderately similar distribution to real ones, but they also fill the real image distribution uncovered by the original dataset, implying their effective DA performance; especially, the CPGGAN-generated images trained without normal brain images distribute wider than the center-concentrating images trained with the normal brain images. Meanwhile, real/synthetic whole brain images clearly distribute differently, due to the real

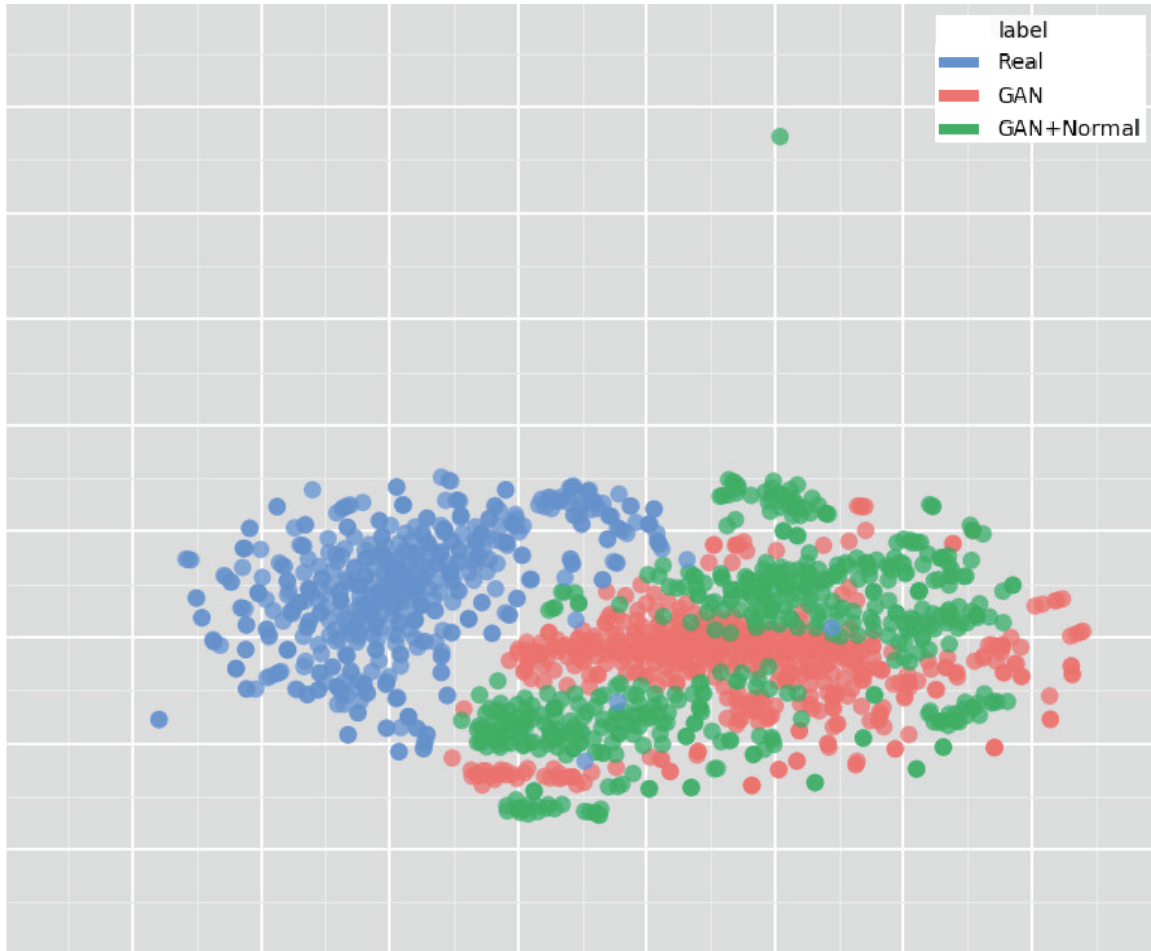


Figure 6-7: T-SNE results with 500 256×256 images per each category: (a) Real tumor images; (b), (c) CPGGAN-generated tumor images, trained without/with additional normal brain images.

MR images' strong anatomical consistency (Fig. 6-7). Considering the achieved high DA performance, the tumor (i.e., ROI) realism/diversity matter more than the whole image realism/diversity, since YOLOv3 look at an image patch instead of a whole image, similarly to most other CNN-based object detectors.

6.5 Conclusion

Without relying on an input benign image, 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 sensitiv-

ity in tumor detection—even with small/fragmented training data from multiple MRI scanners and lazy annotation using highly-rough bounding boxes; in the context of intelligent data wrangling, this attributes to the CPGGANs’ good generalization ability to incrementally synthesize conditional whole images with the real image distribution unfilled by the original dataset, improving the training robustness.

We confirm that the realism and diversity of the generated images, judged by three expert physicians *via* Visual Turing Test, do not imply better detection performance; as the t-SNE results show, the CPGGAN-generated images, trained with additional non-tumor normal images, lack diversity probably because the training less focuses on tumors. Moreover, we notice that adding over-sufficient synthetic images leads to more FPs, but not always higher sensitivity, possibly due to the training data imbalance between real and synthetic images; as the t-SNE results reveal, the CPGGAN-generated tumor bonding boxes have a moderately similar—mutually complementary—distribution to the real ones; thus, GAN-overwhelming training images may decrease the necessary influence of the real samples and harm training, rather than providing robustness. Lastly, image-to-image GAN-based DA just moderately facilitates detection with less additional FPs, probably due to the lack of realism. However, further investigations are needed to maximize the effect of the CPGGAN-based medical image augmentation.

For example, we could verify the effect of further realism in return for less diversity by combining ℓ_1 loss with the WGAN-GP loss for GAN training. We can also combine those CPGGAN-generated images, trained without/with additional brain images, similarly to ensemble learning [124]. Lastly, we plan to define a new GAN loss function that directly optimizes the detection results, instead of realism, similarly to the three-player GAN for optimizing classification results [118].

Overall, minimizing expert physicians’ annotation efforts, our novel CPGGAN-based DA approach sheds light on diagnostic and prognostic medical applications, not limited to brain metastases detection; future studies, especially on 3D bounding box detection with highly-rough annotation, are required to extend our promising results. Along with the DA, the CPGGANs has other potential clinical applications in oncol-

ogy: (i) A data anonymization tool to share patients' data outside their institution for training while preserving detection performance. Such a GAN-based application is reported in Shin *et al.* [76]; (ii) A physician training tool to display random synthetic medical images with abnormalities at both common and rare positions/sizes, by training CPGGANs on highly unbalanced medical datasets (i.e., limited pathological and abundant normal samples, respectively). It can help train medical students and radiology trainees despite infrastructural and legal constraints [13].

Chapter 7

GAN-based Medical Image Augmentation for 3D Detection

7.1 Prologue to Fourth Project

7.1.1 Project Publication

- **Synthesizing Diverse Lung Nodules Wherever Massively: 3D Multi-Conditional GAN-based CT Image Augmentation for Object Detection.** C. Han, Y. Kitamura, A. Kudo, A. Ichinose, L. Rundo, Y. Furukawa, K. Umemoto, H. Nakayama, Y. Li, In International Conference on 3D Vision (3DV), Québec City, Canada, pp. 729–737, September 2019.

7.1.2 Context

Prior to this work, no researchers had tackled 3D GANs for general bounding box-based detection whereas 3D Medical Image Analysis can improve diagnosis by capturing anatomical and functional information. Jin *et al.* had used an image-to-image GAN to generate $64 \times 64 \times 64$ CT images of lung nodules including the surrounding tissues by inputting a VOI centered at a lung nodule, but with a central sphere region erased [77]; however, they had targeted annotation-expensive segmentation, instead of the detection, also translating both nodules/surroundings *via* expensive computation.

Without conditioning a noise-to-image GAN with nodule position, Gao *et al.* had generated $40 \times 40 \times 18$ 3D nodule subvolumes only applicable to their subvolume-based detector using binary classification [75]. Unfortunately, no research had focused on multiple GAN conditions for more versatile 3D GANs while lesions vary in position, size, and attenuation.

7.1.3 Contributions

This project’s primary contribution is to propose a novel 3D pathology-aware multi-conditional GAN called 3D MCGAN for improved 3D bounding box-based detection in general; it translates noise boxes into realistic/diverse $32 \times 32 \times 32$ lung nodules placed naturally at desired position, size, and attenuation on CT scans—inputting the noise box with the surrounding tissues has the effect of combining the noise-to-image and image-to-image GANs. The $32 \times 32 \times 32$ nodule-only generation, not translating the $64 \times 64 \times 64$ surroundings, can decrease computational cost. By so doing, our 3D MCGAN-based DA boosts sensitivity in nodule detection under any nodule size/attenuation at fixed FP rates. Moreover, we find that GAN training with ℓ_1 loss could increase synthetic images’ realism, but decrease DA performance. Using proper augmentation ratio (i.e., 1 : 1) could improve the DA performance. Considering the outstanding realism confirmed by physicians, it could perform as a physician training tool to display realistic medical images with desired abnormalities (i.e., position, size, and attenuation).

7.1.4 Recent Developments

According to their arXiv paper, Xu *et al.* have generated realistic/diverse $64 \times 64 \times 64$ CT images of lung nodules combining the image-to-image GAN with gene expression profiles [125].

7.2 Motivation

Accurate CAD, thanks to recent CNNs, can alleviate the risk of overlooking the diagnosis in a clinical environment. Such great success of CNNs, including diabetic eye disease diagnosis [119], primarily derives from large-scale annotated training data to sufficiently cover the real data distribution. However, obtaining and annotating such diverse pathological images are laborious tasks; thus, the massive generation of proper synthetic training images matters for reliable diagnosis. Researchers usually use classical DA techniques, such as geometric/intensity transformations [88, 89]. However, those one-to-one translated images have intrinsically similar appearance and cannot sufficiently cover the real image distribution, causing limited performance improvement; in this regard, thanks to their good generalization ability, GANs [7] can generate realistic but completely new samples using many-to-many mappings for further performance improvement; GANs showed excellent DA performance in computer vision, including 21% performance improvement in eye-gaze estimation [8].

This GAN-based DA trend especially applies to medical imaging, where the biggest problem lies in small and fragmented datasets from various scanners. For performance boost in various 2D medical imaging tasks, some researchers used noise-to-image GANs for classification [12, 16, 15]; others used image-to-image GANs for object detection [20] and segmentation [121]. However, although 3D imaging is spreading in radiology (e.g., CT and MRI), such 3D medical GAN-based DA approaches are limited, and mostly focus on segmentation [76, 77]—3D medical image generation is more challenging than 2D one due to expensive computational cost and strong anatomical consistency. Accordingly, no 3D conditional GAN-based DA approach exists for general bounding box-based 3D object detection, while it can locate disease areas with physicians’ minimum annotation cost, unlike rigorous 3D segmentation. Moreover, since lesions vary in position/size/attenuation, further GAN-based DA performance requires multiple conditions.

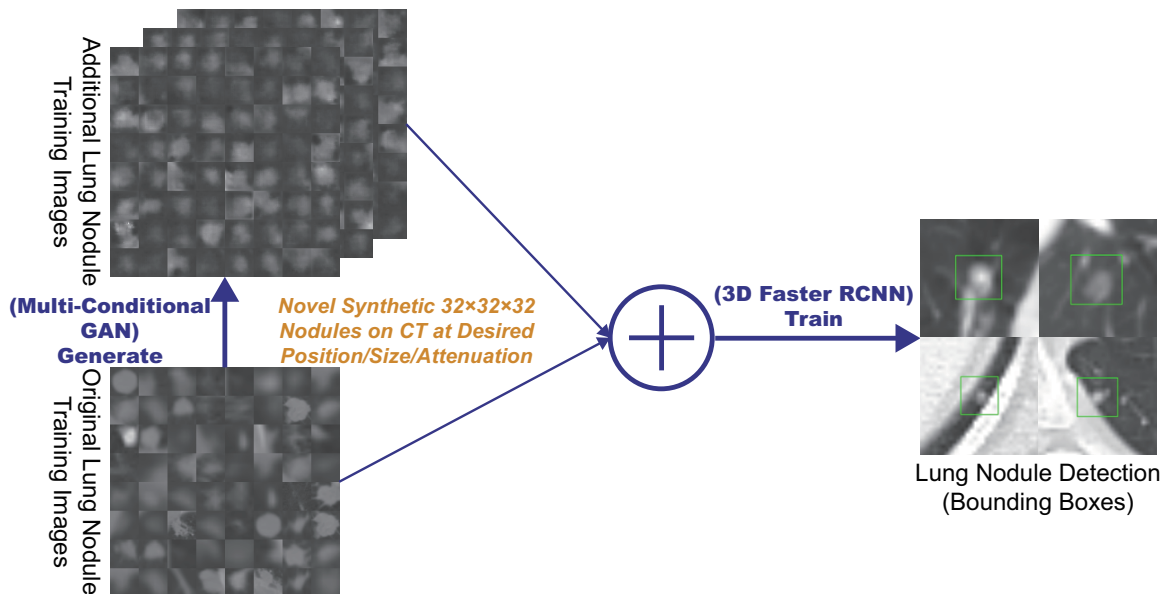


Figure 7-1: 3D MCGAN-based DA for better nodule detection: Our MCGAN generates realistic and diverse nodules naturally on lung CT scans at desired position, size, and attenuation based on bounding boxes, and the CNN-based object detector uses them as additional training data.

So, 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. 7-1), we propose 3D MCGAN to generate $32 \times 32 \times 32$ nodules—such nodule detection is clinically valuable for the early diagnosis/treatment of lung cancer, the deadliest cancer [126]. 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. We also evaluate the synthetic images’ realism *via* Visual Turing Test [91] by two expert physicians, and visualize the data distribution *via* t-SNE [99]. The 3D MCGAN-generated additional training images can achieve higher sensitivity under any nodule size/attenuation at fixed FP rates. Lastly, this study suggests training GANs without ℓ_1 loss and using proper augmentation ratio (i.e., 1 : 1) for better medical GAN-based DA performance.

Research Questions. We mainly address two questions:

- **3D Multiple GAN Conditioning:** How can we condition 3D GANs to naturally place objects of random shape, unlike rigorous segmentation, at desired position/size/attenuation based on bounding box masks?
- **Synthetic Images for DA:** How can we set the number of real/synthetic training data and GAN loss functions to achieve the best detection performance?

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, which even expert physicians cannot distinguish from real ones.
- **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 medical images/annotation.
- **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.

7.3 Materials and Methods

7.3.1 3D MCGAN-based Image Generation

Data Preparation This study exploits the Lung Image Database Consortium image collection (LIDC) dataset [127] containing 1,018 chest CT scans with lung nodules. Since the American College of Radiology recommends lung nodule evaluation using thin-slice CT scans [128], 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 .

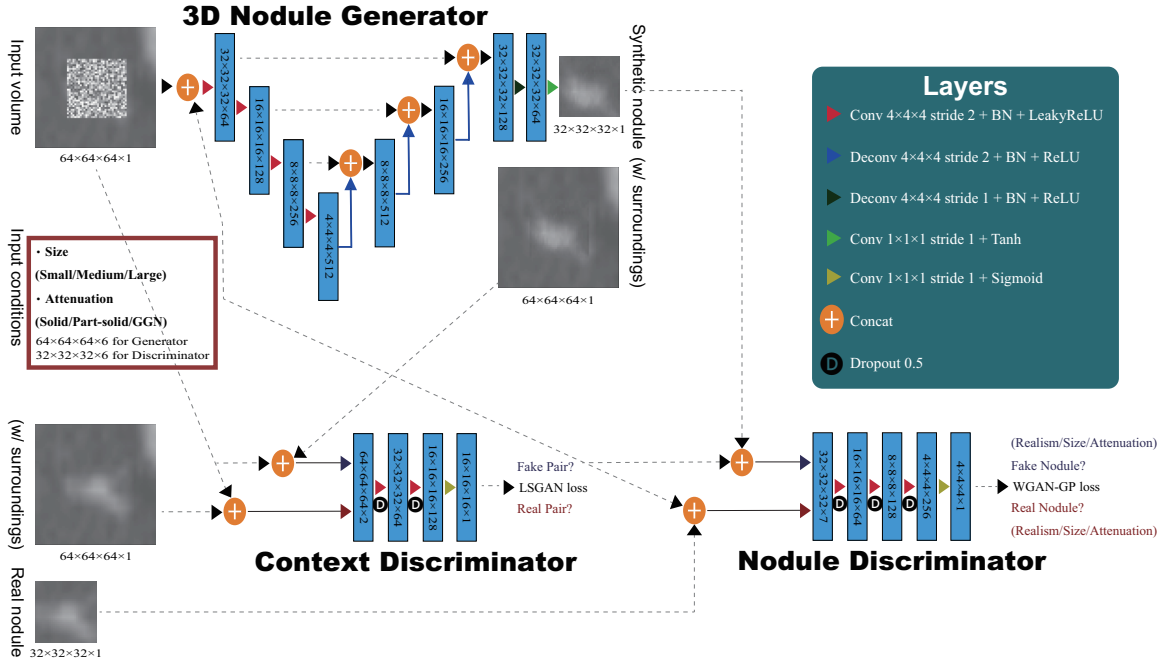


Figure 7-2: Proposed 3D MCGAN architecture for realistic/diverse $32 \times 32 \times 32$ lung CT scan of nodule generation: the context discriminator learns to classify real *vs* synthetic nodule/surrounding pairs while the nodule discriminator learns to classify real *vs* synthetic nodules.

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 ($10 \text{ mm} \leq \text{slice thickness} \leq 20$ 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. The training set contains more average nodules since we exclude patients with too many nodules for the validation/test sets; we arrange a clinical environment-like situation, where we could find more healthy patients than highly diseased ones to conduct anomaly detection.

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. 7-2). 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 (e.g., if the size is small, each voxel of the small condition is filled with 1, while the medium/large condition voxels are filled with 0 to consider the effect of scaling factor). 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 [129]-like generator adopts 4 convolutional layers in encoders and 4 deconvolutional layers in decoders respectively with skip connections to effectively capture both nodule/context information.

We adopt two *Pix2Pix* GAN [42]-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) [105]; the nodule discriminator attempts to classify real *vs* synthetic nodules with size/attenuation conditions using WGAN-GP [39]. 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; similar results are also reported by this work [130] for 2D pedestrian detection without label conditioning. We apply dropout to inject randomness and balance the generator/discriminators. Batch normalization is applied to both convolution (using LeakyReLU) and deconvolution (using ReLU).

Most GAN-based DA approaches use reconstruction ℓ_1 loss [75] to generate realistic images, even modifying it for further realism [77]. However, no one has ever validated whether it really helps DA—it assures synthetic images resembling the original ones, sacrificing diversity; thus, to confirm its influence during classifier training,

we compare our MCGAN objective without/with it, respectively:

$$\begin{aligned}
 G^* &= \arg \min_G \max_{D1, D2} \mathcal{L}_{\text{LSGANs}}(G, D1) \\
 &+ \mathcal{L}_{\text{WGAN-GP}}(G, D2),
 \end{aligned} \tag{7.1}$$

$$\begin{aligned}
 G^* &= \arg \min_G \max_{D1, D2} \mathcal{L}_{\text{LSGANs}}(G, D1) \\
 &+ \mathcal{L}_{\text{WGAN-GP}}(G, D2) + 100\mathcal{L}_{\ell_1}(G).
 \end{aligned} \tag{7.2}$$

We set 100 as a weight for the ℓ_1 loss, since empirically it works well for reducing visual artifacts introduced by the GAN loss and most GAN works adopt the weight [42, 130].

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. As post-processing, we blend bounding boxes’ 3 nearest surfaces from all the boundaries by averaging the values of 6 nearest voxels/itself for 5 iterations. We resample the resulting nodules to their original resolution and map back onto the original CT scans to prepare additional training data.

7.3.2 3D Faster RCNN-based Lung Nodule Detection

3D Faster RCNN is a 3D version of Faster RCNN [123] using multi-task loss with a 27-layer Region Proposal Network of 3D convolutional/batch normalization/ReLU layers. 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 $1 \times 2 \times 3 \times$ MCGAN-based DA (i.e., 632/1, 264/1, 896 additional synthetic training images), (v), (vi), (vii) with $1 \times 2 \times 3 \times$ MCGAN-based DA trained with ℓ_1 loss. During training, we shuffle the real/synthetic image order. We evaluate the detection performance as follows: (i) Free Receiver Operation Characteristic (FROC)

analysis, sensitivity as a function of FPs per scan; (ii) Competition Performance Metric (CPM) score [131], average sensitivity at seven pre-defined FP rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan—this quantifies if a CAD system can identify a significant percentage of 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 with momentum. The input volume size to the network 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.

7.3.3 Clinical Validation *via* Visual Turing Test

To quantitatively evaluate the realism of MCGAN-generated images, we supply, in random order, to two expert physicians a random selection of 50 real and 50 synthetic lung nodule images with all of 2D axial/coronal/sagittal views at the center. They take four classification tests in ascending order: Test1, 2: real *vs* MCGAN-generated $32 \times 32 \times 32$ nodules, trained without/with ℓ_1 loss; Test3, 4: real *vs* MCGAN-generated $64 \times 64 \times 64$ nodules with surroundings without/with ℓ_1 loss.

7.3.4 Visualization *via* t-SNE

To visually analyze the distribution of real/synthetic images, we use t-SNE [99] on a random selection of 500 real, 500 synthetic, and 500 ℓ_1 loss-added synthetic nodule images, with a perplexity of 100 for 1,000 iterations to get a 2D representation. We normalize the input images to $[0, 1]$.

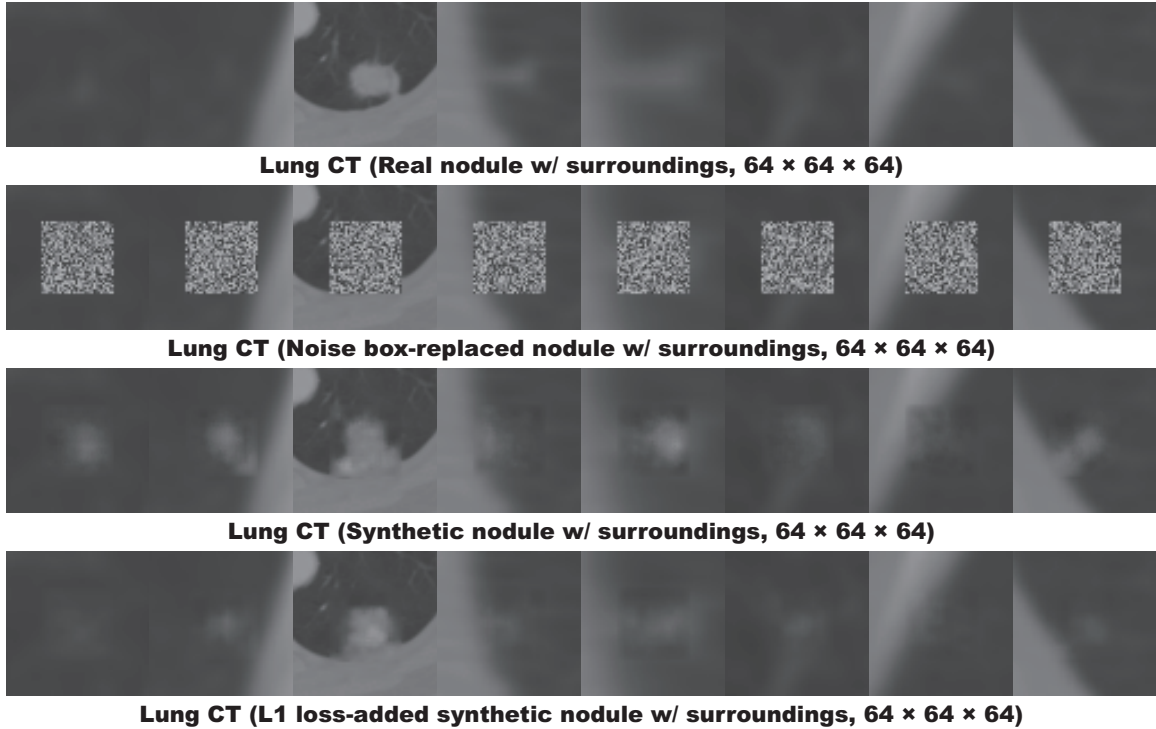


Figure 7-3: 2D axial view of example real/synthetic $64 \times 64 \times 64$ CT scans of lung nodules with surrounding tissues; 3D MCGANs generate only $32 \times 32 \times 32$ nodules.

7.4 Results

7.4.1 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. 7-3). Especially, when trained without ℓ_1 loss, those synthetic nodules look clearly more different from the original real ones, including slight shading difference.

Table 7.1: 3D Faster RCNN nodule detection results (CPM) of seven DA setups (IoU \geq 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 (%)	CPM by Size (%)			CPM by Attenuation (%)		
		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 \times 3D MCGAN-based DA	55.0	45.2	68.3	66.2	69.9	52.1	24.4
+ 2 \times 3D MCGAN-based DA	52.7	44.7	67.4	42.9	65.5	40.7	28.9
+ 3 \times 3D MCGAN-based DA	51.2	41.1	64.4	66.2	61.6	57.9	27.7
+ 1 \times 3D MCGAN-based DA w/ ℓ_1	50.8	43.0	63.3	55.6	62.6	47.1	27.1
+ 2 \times 3D MCGAN-based DA w/ ℓ_1	50.9	40.6	64.4	65.4	64.9	43.6	23.3
+ 3 \times 3D MCGAN-based DA w/ ℓ_1	47.9	38.9	59.4	61.7	59.6	50.7	22.6

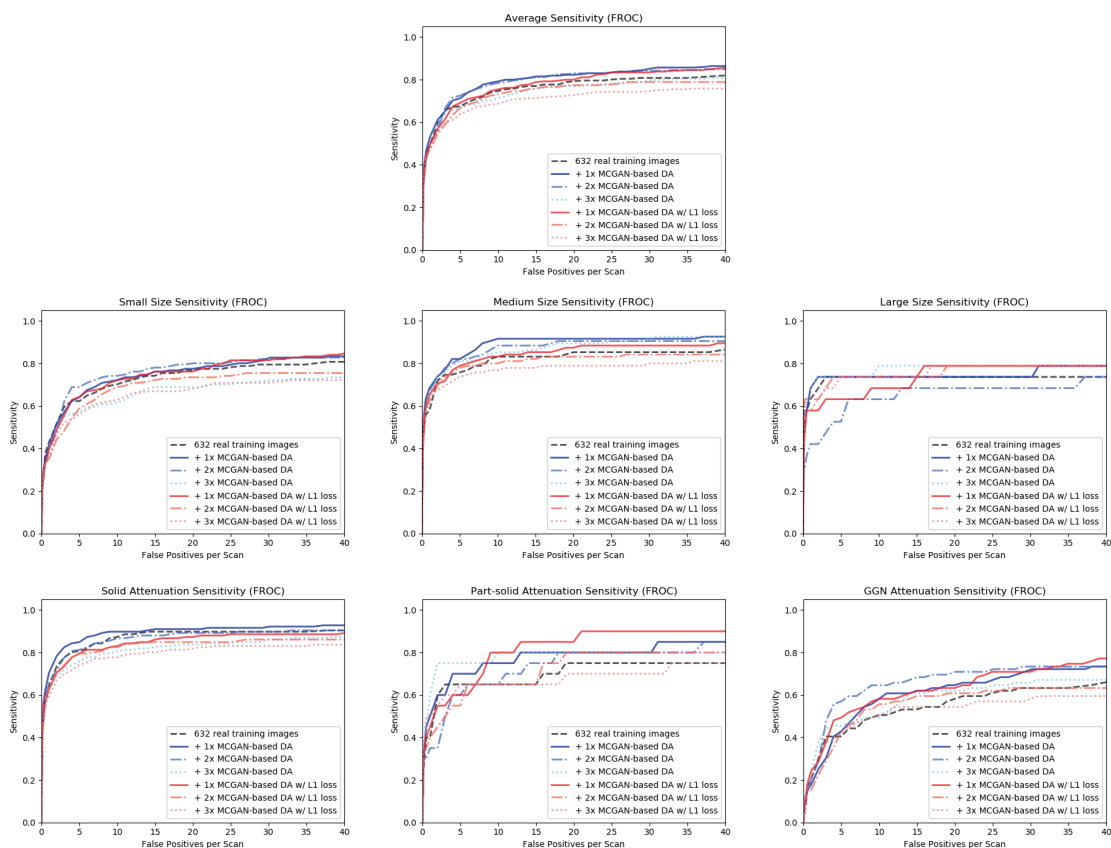


Figure 7-4: FROC curves of seven DA setups by average/size/attenuation.

7.4.2 Lung Nodule Detection Results

Table 7.1 and Fig. 7-4 show 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 \times MCGAN-based DA without ℓ_1 loss outperforms training only with real images

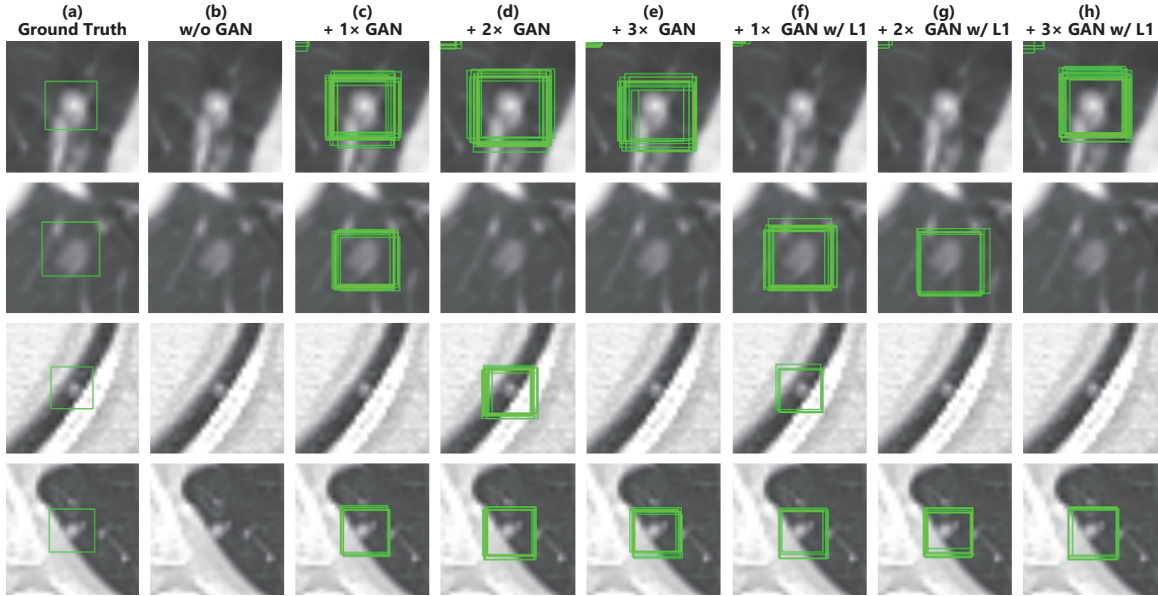


Figure 7-5: Example detection results of seven DA setups on four different images, compared against the ground truth (detection threshold 0.5): (a) ground truth; (b) without GAN-based DA; (c), (d), (e) with $1\times/2\times/3\times$ 3D MCGAN-based DA; (f), (g), (h) with $1\times/2\times/3\times$ ℓ_1 loss-added 3D MCGAN-based DA.

under any size/attenuation in terms of CPM, achieving average CPM improvement by 0.032. It especially boosts nodule detection performance with larger size and lower attenuation. Fig. 7-5 visually reveals its ability to alleviate the risk of overlooking the nodule diagnosis with clinically acceptable FPs (i.e., the highly-overlapping bounding boxes around nodules only require a physician’s single check by switching on/off transparent alpha-blended annotation on CT scans). 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 7.2: Visual Turing Test results by two physicians for classifying 50 real *vs* 50 3D MCGAN-generated images: Test1, 2: $32 \times 32 \times 32$ lung nodules, trained without/with ℓ_1 loss; Test3, 4: $64 \times 64 \times 64$ nodules with surrounding tissues, trained without/with ℓ_1 loss. Accuracy denotes the physicians’ successful classification ratio between the real/synthetic images.

		Accuracy (%)	Real as Real (%)	Real as Synt (%)	Synt as Real (%)	Synt as Synt (%)
Test 1	Physician 1	43	38	62	52	48
	Physician 2	43	26	74	40	60
Test 2	Physician 1	57	44	56	30	70
	Physician 2	53	22	78	16	84
Test 3	Physician 1	62	50	50	26	74
	Physician 2	79	64	36	6	94
Test 4	Physician 1	58	42	58	26	74
	Physician 2	66	72	28	40	60

7.4.3 Visual Turing Test Results

As Table 7.2 shows, expert physicians fail to classify real *vs* MCGAN-generated nodules without surrounding tissues—even regarding the synthetic nodules trained without ℓ_1 loss more realistic than the real ones. Contrarily, they relatively recognize the synthetic nodules with surroundings due to slight shading difference between the nodules/surroundings, especially when trained without the reconstruction ℓ_1 loss. Considering the synthetic images’ realism, CPGGANs might perform as a tool to train medical students and radiology trainees when enough medical images are unavailable, such as abnormalities at rare position/size/attenuation. Such GAN applications are clinically promising [13].

7.4.4 T-SNE Results

Implying their effective DA performance, synthetic nodules have a similar distribution to real ones, but concentrated in left inner areas with less real ones especially when trained without ℓ_1 loss (Fig. 7-6)—using only GAN loss during training can avoid overwhelming influence from the real image samples, resulting in a moderately similar distribution; thus, those synthetic images can partially fill the real image distribution uncovered by the original dataset.

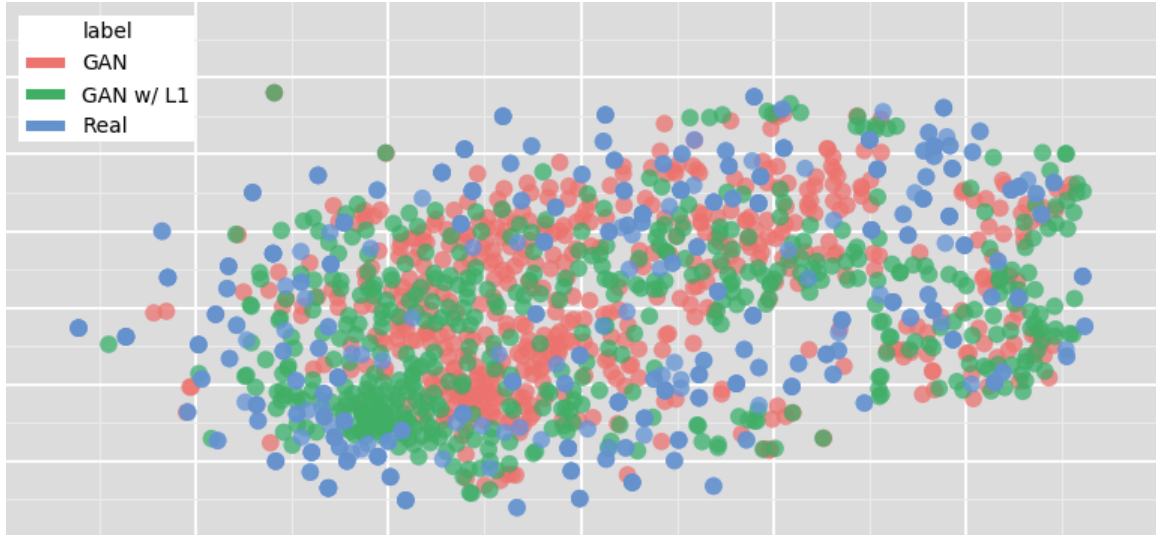


Figure 7-6: T-SNE plot with 500 $32 \times 32 \times 32$ nodule images per each category: (a), (b) 3D MCGAN-generated nodules, trained without/with ℓ_1 loss; (c) real nodules.

7.5 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; they allow to cover the real image distribution unfilled by the original dataset, improving the training robustness.

Surprisingly, we find that adding over-sufficient synthetic images produces worse results due to the real/synthetic image balance; as t-SNE results show, the synthetic images only partially cover the real image distribution, and thus GAN-overwhelming training images rather harm training. Moreover, we notice that GAN training without ℓ_1 loss obtains better DA performance thanks to increased diversity providing robustness; also, expert physicians confirm their sufficient realism without ℓ_1 loss.

Overall, our 3D MCGAN could help minimize expert physicians’ time-consuming annotation tasks and overcome the general medical data paucity, not limited to lung

CT nodules. As future work, we will investigate the MCGAN-based DA results without size/attenuation conditioning to confirm their influence on DA performance. Moreover, we will compare our DA results with other non-GAN-based recent DA approaches, such as Mixup [50] and Cutout [52]. For further performance boost, we plan to directly optimize the detection results for MCGANs, instead of realism, similarly to the three-player GAN for classification [118]. Lastly, we will investigate how our MCGAN can perform as a physician training tool to display random realistic medical images with desired abnormalities (i.e., position/size/attenuation conditions) to help train medical students and radiology trainees despite infrastructural and legal constraints [13].

Chapter 8

Discussions on Developing Clinically Relevant AI-Powered Diagnosis Systems

8.1 Feedback from Physicians

8.1.1 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 the following questionnaire within 2 weeks from December 6th, 2019 after reading 10 summary slides written in Japanese¹ about general 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 [132]) evaluation: Likert scale 1 = very negative, 2 = negative, 3 = neutral, 4 =

¹Available via Dropbox: <https://www.dropbox.com/sh/bacowc3ilz1p1r3/AAENS9SyjArHq8BntgaODLb2a?dl=0>

positive, 5 = very positive.

- **Question 1:** Are you keen to exploit medical AI in general when it achieves accurate and reliable performance in the near future? (five-point Likert scale) Please tell us your expectations, wishes, and worries (free comments).
- **Question 2:** What do you think about using GAN-generated images for DA? (five-point Likert scale) Please tell us your expectations, wishes, and worries (free comments).
- **Question 3:** What do you think about using GAN-generated images for physician training? (five-point Likert scale) Please tell us your expectations, wishes, and worries (free comments).
- **Question 4:** Any comments or suggestions about our projects towards developing clinically relevant AI-powered systems based on your daily diagnosis experience?

8.1.2 Results

We show the questions and Japanese physicians' corresponding answers.

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

- **Likert scale** Project-related physicians: 5 5 5 (average: 5)
Project non-related radiologists: 5 5 3 4 5 5 (average: 4.5)
- **Free comments** (one comment for each physician)
- As radiologists, we need AI-based diagnosis during image interpretation as soon as possible.
- It is common to conduct further medical examinations when identifying disease is difficult from CT/MR images; thus, if AI-based diagnosis outperforms that of physicians, such clinical decision support systems could prevent unnecessary

examinations. Moreover, recently lung cancer misdiagnosis occurred in Japan, but AI technologies may prevent such death caused by misdiagnosis.

- The lack of diagnosticians is very evident in Healthcare, so AI has great potential to support us. It may be already applicable without severe problems for typical disease cases.
- I am looking forward to its practical applications, especially at low or zero price.
- I would like to use AI-based diagnosis as a kind of data, but it is yet uncertain how much I trust AI.
- I am wondering whether such systems will become popular due to practical problems such as introduction cost.
- The definition of *accurate and reliable* is unclear. Since a physician's annotation is always subjective, we cannot claim that AI-based diagnosis is really correct even if AI diagnoses similarly to the specific physician. Because I do not believe other physicians' diagnosis, but my own eyes, I would use AI just to identify abnormal candidates.

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).

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

- **Likert scale** Project-related physicians: 5 5 4 (average: 4.7)
Project non-related radiologists: 4 5 4 4 4 4 (average: 4.2)
- **Free comments** (one comment for each physician)
- Achieved accuracy improvement shows its superiority in identifying diverse disease.
- It would be effective, especially as rare disease training data.
- I am looking forward to the future with advanced GAN technology.
- It significantly improves detection sensitivity; but I am also curious about its influence on other metrics, such as specificity.
- If Deep Learning could be more effective, we should introduce it; but anonymization would be important for privacy preservation.
- Achieved accuracy improvement shows its superiority in identifying diverse disease.
- It would be effective to train AI on data-limited disease, but which means that AI is inferior to humans.
- It would be helpful if such DA improves accuracy and reliability. Since I am not familiar with AI and a generator/classifier's failure judgment mechanisms, I am uncertain whether it will really increase reliability though.

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 are 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?

- **Likert scale** Project-related physicians: 3 4 3 (average: 3.3)
Project non-related radiologists: 3 5 2 3 2 3 (average: 3)
- **Free comments** (one comment for each physician)
- In future medical care, physicians should actively introduce and learn new technology; in this sense, GAN technology should be actively used for physician training in rare diseases.
- It could be useful for medical student training, which aims for 85% accuracy by covering typical cases. But expert physician training aims for over 85% accuracy by comparing typical/atypical cases and acquiring new understanding—real atypical images are essential.
- In physician training, we use radiological images after definite diagnosis, such as pathological examination—but, we actually lack rare disease cases. Since the GAN-generated images' realism fluctuates based on image augmentation schemes and available training images, further realistic image generation of the rare cases would help the physician training.
- It depends on how to construct the system.
- Which specific usage is assumed for such physician training?
- I cannot state an opinion before actually using the system, but I strongly recognize the importance of looking at real images.
- I do not exactly understand in which situation such physician training is used, but eventually training with realistic images would be also helpful. However, if real images are available, using them would be better.

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 with GAN-generated atypical images, along with pathology-aware GAN-based extrapolation, further GAN-based extrapolation would be valuable.

Question 4: Any comments or suggestions about our pathology-aware GAN projects towards developing clinically relevant AI-powered systems based on your daily diagnosis experience?

- This approach will change the way physicians work. I have high expectations for AI-based diagnosis, so I hope it to overcome the legal barrier.
- For now, please show small abnormal findings, such as nodules and ground glass opacities—it would halve radiologists' efforts. Then, we could develop accurate diagnosis step by step.
- Showing abnormal findings with their shapes/sizes/disease names would increase diagnosis accuracy. But I also would like to know how diagnosticians' roles change after all.
- I hope that this approach will lead to physicians' work reduction in the future.
- Please develop reliable AI systems by increasing accuracy with the GAN-based image augmentation.
- GANs can generate typical images, but not atypical images; this would be the next challenge.
- AI can alert physicians to detect typical cases, and thus decrease interpretation time; however, it may lead to the diagnosticians' easy diagnosis without much consideration. Especially in Japan, we currently often conduct unnecessary diagnostic tests, so the diagnosticians should be more responsible of their own duties after introducing AI.

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 clinician-friendly systems with new practice guidelines, and overcoming legal/financial constraints.

8.2 AI and Healthcare Workshop

8.2.1 Methods for Workshop Evaluation

Convolutional Neural Networks (CNNs) have achieved accurate and reliable Computer-Aided Diagnosis (CAD), occasionally outperforming expert physicians [1, 133, 134]. However, such research results cannot be easily applied to a clinical environment: 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:** As its program shows (Table 8.1), 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. Supported by GCL

Table 8.1: Workshop program to *i*) know the overview of Medical Image Analysis and *ii*) find the intrinsic gap and its solutions between AI researchers and Healthcare workers. * indicates activities given by a facilitator (i.e., the first author), such as lectures.

Time (mins)	Activity
	Introduction
10	1. Explanation of the workshop’s purpose and flow*
10	2. Self-introduction and explanation of motivation for participation
5	3. Grouping into two groups based on background*
	Learning: Knowing Medical Image Analysis
15	1. TED speech video watching: <i>Artificial Intelligence Can Change the future of Medical Diagnosis</i> *
35	2. Lecture: Overview of Medical Image Analysis including state-of-the-art research, well-known challenges/solutions, and our pathology-aware GAN projects summary* (its video in Japanese: https://youtu.be/rTQLknPvnqs)
10	3. Sharing expectations, wishes, and worries about Medical Image Analysis (its video in Japanese: https://youtu.be/ILPEGga-hkY)
10	Intermission
	Thinking: Finding How to Develop Robust Medical AI
25	1. Identifying the intrinsic gap between AI/Healthcare sides after sharing their common and different thinking/working styles
60	2. Finding how to develop gap-bridging medical AI fitting into a clinical environment in five years
10	Intermission
	Summary
25	1. Presentation
10	2. Sharing workshop impressions and ideas to apply obtained knowledge (its video in Japanese: https://youtu.be/F31tPR3m8hs)
5	3. Answering a questionnaire about satisfaction/further comments
5	4. Closing remarks*

program, this workshop was held on March 17th, 2019 at Nakayama Future Factory, Open Studio, The University of Tokyo, Tokyo, Japan.

8.2.2 Results

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

Gap Between AI and Healthcare Sides

Gap 1: AI, including Deep Learning, does not provide clear decision criteria, does it make physicians reluctant to use it in a clinical environment, especially for diagnosis?

- **Healthcare side:** We rather expect applications other than diagnosis. If we use AI for diagnosis, instead of replacing physicians, we suppose a *reliable second opinion*, such as alert to avoid misdiagnosis, based on various clinical data not limited to images—every single diagnostician is anxious about their diagnosis. AI only provides minimum explanation, such as a heatmap showing attention, which makes persuading not only the physicians but also patients difficult; so, the physicians’ intervention is essential for intuitive explanation. Methodological safety and feeling safe are different. In this sense, pursuing explainable AI generally decreases AI’s diagnostic accuracy [135], so physicians should still serve as mediators by engaging in high-level conversation or interaction with patients. Moreover, according to the medical law in most countries including Japan, only doctors can make the final decision. The first autonomous AI-based diagnosis without a physician was cleared by the Food and Drug Administration in the US in 2018 [136], but such a case is exceptional.
- **AI side:** Compared with other systems or physicians, Deep Learning’s explanation is not particularly poor, so we require too severe standards for AI; the word *AI* is excessively promoting anxiety and perfection. If we could thoroughly verify the reliability of its diagnosis against physicians by exploring uncertainty measures [137], such intuitive explanation would be optional.

Gap 2: Are there any benefits to actually introducing medical AI?

- **Healthcare side:** After all, even if AI can achieve high accuracy and convenient operation, hospitals would not introduce it without any commercial benefits. Moreover, small clinics, where physicians are desperately needed, often do not have CT or MRI scanners [138].

- **AI side:** The commercial deployment of medical AI is strongly tied to diagnostic accuracy [139]; so, if it can achieve significantly outstanding accuracy at various tasks in the near future, patients would not visit hospitals/clinics without AI. Accordingly, introducing medical AI would become profitable in five years.

Gap 3: Is medical AI’s diagnostic accuracy reliable?

- **Healthcare side:** To evaluate AI’s diagnostic performance, we should consider many metrics, such as sensitivity and specificity. Moreover, its generalization ability for medical data highly relies on inter-scanner/inter-individual variability [140]. How can we evaluate whether it is suitable as a clinically applicable system?
- **AI side:** Generally, alleviating the risk of overlooking the diagnosis is the most important, so sensitivity matters more than specificity unless their balance is highly disturbed. Recently, such research on medical AI that is robust to different datasets is active [96].

How to Develop Medical AI Fitting into a Clinical Environment in Five Years

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 [141] (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 [142]. It

could be also applied to segmentation for radiation therapy [143], neurosurgery navigation [144], and pressure ulcers' echo evaluation. Along with improving the diagnosis, it would also make the physicians' workflow easier, such as by denoising [145]. 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 [146]. 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 [147] 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 than to do it in Japan alone. 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 [148]. To handle changes in diagnostic criteria/equipment and overcome dataset/task dependency, it is necessary to establish a common database creation workflow [149] by regularly entering electronic medical records into the database. For reducing

data acquisition/annotation cost, AI techniques, such as GAN-based DA [21] and domain adaptation [150], 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 [151] operated with less staff. Medical manufacturers could also standardize data format [152], 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 [153], so they could help rural areas and developing countries, where physician shortage is severe [138], at relatively low cost.

How: Safety and feeling safe

- **Challenges:** Considering multiple metrics, such as sensitivity and specificity [154], and dataset/task dependency [155], accuracy could be unreliable, so ensuring safety is challenging. Moreover, reassuring physicians and patients is important to actually use AI in a clinical environment [156].
- **Solutions:** We should integrate various clinical data, such as blood test biomarkers and multiomics, with images [142]. Moreover, developing bias-robust technology is important since confounding factors are inevitable [157]. To prevent oversight, prioritizing sensitivity over specificity is essential while maintaining a balance [158]. We should also devise education for medical AI users, such as result interpretation, to reassure patients [159].

Chapter 9

Conclusion

9.1 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 [11]. 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 [96, 160, 161]. Towards clinically relevant implementation for the DA and physician training, we find effective loss functions and training schemes for each of them [15, 16]—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.

Specifically, our results imply that GAN training without ℓ_1 loss, using proper augmentation ratio (i.e., 1 : 1), and further refining synthetic images' texture/shape could improve the DA performance, whereas discarding weird-looking synthetic images to humans is unnecessary; for example, adding over-sufficient GAN-generated training images leads to more FPs in detection, but not always higher sensitivity, due to the real/synthetic training data balance (both of their distributions are biased, but differently). Regarding the physician training, GAN training with ℓ_1 loss, GAN

training on additional normal images, and post-processing, such as by Poisson image editing [162], could improve the synthetic images’ realism; for instance, the GAN training on normal images along with pathological ones, remarkably facilitates the realism of both healthy and pathological parts while they do not include abnormality.

Because such excellent realism and diversity can be achieved by GAN-based interpolation and extrapolation, we propose novel 2D/3D pathology-aware GANs for bounding box-based pathology detection [20, 21]: (*Interpolation*) The GAN-based medical image augmentation is reliable because medical modalities (e.g., X-ray, CT, MRI) can display the human body’s strong anatomical consistency at fixed position while clearly reflecting inter-subject variability [9, 10]—this is different from natural images, where various objects can appear at any position; (*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 [19].

After conducting a questionnaire survey about our GAN projects for 9 physicians and holding a workshop about how to develop medical AI fitting into a clinical environment for 7 professionals with various AI and/or Healthcare background, we confirm our pathology-aware GANs’ clinical relevance for diagnosis: (*DA*) They could be integrated into a clinical decision support system; since CT has a much higher earning rate and longer interpretation time than MRI (16% to 3% and 2-20 minutes to 1 minute in Japan), alerting abnormal findings on CT, such as nodules/ground glass opacities, would halve radiologists’ efforts and increase hospitals’ financial outcomes; (*Physician training*) They could perform as a non-expert physician training tool; when the normal training images are sufficiently available, we can stably generate typical pathological images useful for medical student training, thanks to the excellent interpolation; but it is still challenging to generate atypical images needed for expert physician training. Whereas our pathology-aware bounding box conditioning largely improves extrapolation ability, better DA and physician training would require further GAN-based extrapolation.

9.2 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 [163], (ii) generation with both image and gene expression conditions [125], and (iii) transfer learning among different body parts and disease types [164]. 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 CO-ordinate GAN (COCO-GAN) to generate state-of-the-art realistic and seamless full images [163]. 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 and physician training.

We can also condition the GANs both on the image features and gene expression profiles to non-invasively identify molecular properties of disease. By so doing, Xu *et al.* succeeded to produce $60 \times 60 \times 60$ realistic synthetic CT images of lung nodules [125]. If the gene expression data are available, such condition fusing could be helpful for the medical DA and physician training.

Such information conversion, not limited to the GAN conditioning, should locate in the core of future Medical Image Analysis to overcome the data paucity. Whereas the transfer learning from large-scale natural image/video datasets for CNNs is already common in Medical Image Analysis, such pre-trained models cannot extract general human anatomical features. Accordingly, pre-training on large-scale 3D medical volumes for CNNs, such as CT and MRI, significantly outperformed the pre-training on natural videos or training from scratch for classification and segmenta-

tion [164] both by accuracy and training convergence speed. Similarly, transfer learning from mammography for the CNNs also significantly improved mass detection on digital breast tomosynthesis slices [165]. Such transfer learning across different body parts and disease types for the GANs would also largely improve their extrapolation ability.

Bibliography

- [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] J. Hsieh, *Computed tomography: Principles, design, artifacts, and recent advances*. SPIE Bellingham, WA, 2009.
- [10] R. W. Brown, Y. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: Physical principles and sequence design*. John Wiley & Sons, 2014.

- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.

- [22] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. De Jong, *et al.*, “Radiomics: The bridge between medical imaging and personalized medicine,” *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, p. 749, 2017.
- [23] D. Rueckert, B. Glocker, and B. Kainz, “Learning clinically useful information from images: Past, present and future,” *Med. Image Anal.*, vol. 33, p. 13, 18.
- [24] J. Yap, W. Yolland, and P. Tschandl, “Multimodal skin lesion classification using deep learning,” *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, 2018.
- [25] J. Lao, Y. Chen, Z. Li, Q. Li, J. Zhang, *et al.*, “A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme,” *Sci. Rep.*, vol. 7, no. 1, p. 10353, 2017.
- [26] 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.
- [27] E. Yiannakopoulou, N. Nikiteas, D. Perrea, and C. Tsigris, “Virtual reality simulators and training in laparoscopic surgery,” *Int. J. Surg.*, vol. 13, pp. 60–64, 2015.
- [28] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, and M. Mazurowski, “Deep learning for identifying radiogenomic associations in breast cancer,” *Comput. Biol. Med.*, vol. 109, pp. 85–90, 2019.
- [29] L. L. Geyer, U. J. Schoepf, F. G. Meinel, J. W. Nance Jr, G. Bastarrika, *et al.*, “State of the art: Iterative CT reconstruction techniques,” *Radiology*, vol. 276, no. 2, pp. 339–357, 2015.
- [30] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from picture to proton*. Cambridge university press, 2017.
- [31] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [32] Y. Bengio and G. Marcus, “DEBATE: Yoshua Bengio and Gary Marcus on the best way forward for AI.” Debate, 2019.
- [33] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *J. Physiol.*, vol. 148, no. 3, pp. 574–591, 1959.
- [34] 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.
- [35] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. International Conference on Learning Representations (ICLR)*, *arXiv preprint arXiv:1312.6114*, 2014.

- [36] J. 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.
- [37] 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.
- [38] 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.
- [39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” *arXiv preprint arXiv:1704.00028*, 2017.
- [40] 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.
- [41] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, *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.
- [42] 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.
- [43] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning what and where to draw,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 217–225, 2016.
- [44] H. Park, Y. Yoo, and N. Kwak, “MC-GAN: Multi-conditional generative adversarial network for image synthesis,” in *Proc. British Machine Vision Conference (BMVC)* *arXiv preprint arXiv:1805.01123*, 2018.
- [45] K. Hung, C. C. Lee, and S. Choy, “Ubiquitous health monitoring: Integration of wearable sensors, novel sensing techniques, and body sensor networks,” in *Mobile Health*, pp. 319–342, Springer, 2015.
- [46] M. J. Yaffe, “Emergence of big data and its potential and current limitations in medical imaging,” *Semin. Nucl. Med.*, vol. 49, no. 2, pp. 94–104, 2019.
- [47] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” *arXiv preprint arXiv:1712.04621*, 2017.
- [48] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

- [49] D. J. Heeger, “The representation of visual stimuli in primary visual cortex,” *Curr. Dir. Psychol. Sci.*, vol. 3, no. 5, pp. 159–163, 1994.
- [50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [51] Y. Tokozume, Y. Ushiku, and T. Harada, “Between-class learning for image classification,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5486–5494, 2018.
- [52] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [53] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *Proc. International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.
- [54] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation policies from data,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019.
- [55] A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” *arXiv preprint arXiv:1711.04340*, 2017.
- [56] C. Han, L. Rundo, K. Murao, Z. Á. Milacski, *et al.*, “GAN-based multiple adjacent brain MRI slice reconstruction for unsupervised Alzheimer’s disease diagnosis,” in *Proc. International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, 2019. In press.
- [57] 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.
- [58] T. Park and G. Casella, “The bayesian Lasso,” *J. Am. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008.
- [59] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. Royal Stat. Soc. Ser. B Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.
- [60] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
- [61] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *J. Jap. Soc. Art.*, vol. 14, no. 771-780, p. 1612, 1999.
- [62] R. Hou, D. Zhou, R. Nie, D. Liu, and X. Ruan, “Brain CT and MRI medical image fusion using convolutional neural networks and a dual-channel spiking cortical model,” *Med. Biol. Eng. Comput.*, vol. 57, no. 4, pp. 887–900, 2019.

- [63] H. Chen, Y. Qi, Y. Yin, T. Li, X. Liu, *et al.*, “Mmfnet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma,” *arXiv preprint arXiv:1812.10033*, 2018.
- [64] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, *et al.*, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Trans. Med. imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [65] J. Zbontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” *J. Mach. Learn. Res.*, vol. 17, no. 1-32, p. 2, 2016.
- [66] B. Romera-Paredes and P. Torr, “An embarrassingly simple approach to zero-shot learning,” in *Proc. International Conference on Machine Learning (ICML)*, pp. 2152–2161, 2015.
- [67] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing machines,” *arXiv preprint arXiv:1410.5401*, 2014.
- [68] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proc. International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.
- [69] 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.
- [70] 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.
- [71] T. Malygina, E. Ercheva, 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.
- [72] 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.
- [73] F. Konidaris, T. Tagaris, M. Sdraka, and A. Stafylopatis, “Generative adversarial networks as an advanced data augmentation technique for MRI data,” in *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pp. 48–59, 2019.

- [74] 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.
- [75] 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.
- [76] 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 *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, pp. 1–11, 2018.
- [77] 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.
- [78] S. Wang, “Competencies and experiential requirements in radiology training,” in *Radiology Education*, pp. 55–66, Springer, 2012.
- [79] 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.
- [80] 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.
- [81] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina, “Biomedical data augmentation using generative adversarial neural networks,” in *Proc. International Conference on Artificial Neural Networks (ICANN)*, pp. 626–634, 2017.
- [82] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proc. International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.
- [83] G. Kwon, C. Han, and D.-s. Kim, “Generation of 3D brain MRI using auto-encoding generative adversarial networks,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 118–126, 2019.
- [84] L. Rundo, C. Militello, G. Russo, S. Vitabile, M. C. Gilardi, and G. Mauri, “GTVcut for neuro-radiosurgery treatment planning: An MRI brain cancer seeded image segmentation method based on a cellular automata model,” *Nat. Comput.*, pp. 1–16, 2017.

- [85] D. Shen, G. Wu, and H. I. Suk, “Deep learning in medical image analysis,” *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.
- [86] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, *et al.*, “Brain tumor segmentation with deep neural networks,” *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.
- [87] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, *et al.*, “Deep learning for health informatics,” *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, 2017.
- [88] 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.
- [89] 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.
- [90] 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.
- [91] 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.
- [92] 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.
- [93] D. Lopez-Paz and M. Oquab, “Revisiting classifier two-sample tests,” in *Proc. International Conference on Learning Representations (ICLR)*, *arXiv preprint arXiv:1610.06545*, 2017.
- [94] 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)*, 2017.
- [95] 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.
- [96] 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.

- [97] J. Ker, L. Wang, J. Rao, and T. Lim, “Deep learning applications in medical image analysis,” *IEEE Access*, vol. 6, pp. 9375–9389, 2017.
- [98] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Med. Image Anal.*, vol. 58, p. 101552, 2019.
- [99] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [100] 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.
- [101] 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.
- [102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.
- [103] 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.
- [104] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. International Conference on Learning Representations (ICLR)*, *arXiv preprint arXiv:1409.1556*, 2015.
- [105] 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.
- [106] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6924–6932, 2017.
- [107] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.
- [108] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proc. International Conference on Computational Statistics (COMPSTAT)*, pp. 177–186, 2010.

- [109] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, “Benchmark analysis of representative deep neural network architectures,” *IEEE Access*, vol. 6, pp. 64270–64277, 2018.
- [110] K. Chang, H. X. Bai, H. Zhou, C. Su, W. L. Bi, *et al.*, “Residual convolutional neural network for the determination of IDH status in low-and high-grade gliomas from MR imaging,” *Clin. Cancer Research*, vol. 24, no. 5, pp. 1073–1081, 2018.
- [111] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, “Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI,” *J. Magn. Reson. Imaging*, vol. 49, no. 4, pp. 939–954, 2019.
- [112] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [113] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Statist.*, pp. 65–70, 1979.
- [114] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “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.
- [115] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, “Densenet: Implementing efficient convnet descriptor pyramids,” *arXiv preprint arXiv:1404.1869*, 2014.
- [116] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1027–1035, 2007.
- [117] 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.
- [118] 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.
- [119] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [120] N. D. Arvold, E. Q. Lee, M. P. Mehta, K. Margolin, B. M. Alexander, *et al.*, “Updates in the management of brain metastases,” *Neuro Oncol.*, vol. 18, no. 8, pp. 1043–1065, 2016.

- [121] O. Bailo, D. Ham, and Y. Min Shin, “Red blood cell image generation for data augmentation using conditional generative adversarial networks,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- [122] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [123] 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.
- [124] T. G. Dietterich, “Ensemble learning,” *The Handbook of Brain Theory and Neural Networks*, vol. 2, pp. 110–125, 2002.
- [125] 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.
- [126] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA-Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, 2019.
- [127] 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.
- [128] A. A. A. Setio, A. Traverso, T. De Bel, M. S. N. Berens, C. van den Bogaard, *et al.*, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge,” *Med. Image Anal.*, vol. 42, pp. 1–13, 2017.
- [129] 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.
- [130] X. Ouyang, Y. Cheng, Y. Jiang, C. L. Li, and P. Zhou, “Pedestrian-synthesis-GAN: Generating pedestrian data in real scene and beyond,” *arXiv preprint arXiv: 1804.02047*, 2018.
- [131] 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.
- [132] I. E. Allen and C. A. Seaman, “Likert scales and data analyses,” *Qual. Prog.*, vol. 40, no. 7, pp. 64–65, 2007.

- [133] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, *et al.*, “Deep neural networks improve radiologists’ performance in breast cancer screening,” in *Proc. International Conference on Medical Imaging with Deep Learning (MIDL)*, *arXiv preprint arXiv:1907.08612*, 2019.
- [134] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.
- [135] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [136] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, “Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices,” *NPJ Digit. Med.*, vol. 1, no. 1, p. 39, 2018.
- [137] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, “Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation,” *Med. Image Anal.*, vol. 59, p. 101557, 2020.
- [138] G. R. Jankharia, “Commentary-radiology in India: The next decade,” *Indian J. Radiol. Imaging*, vol. 18, no. 3, p. 189, 2008.
- [139] S. Vollmer, B. A. Mateen, G. Bohner, F. J. Király, R. Ghani, *et al.*, “Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness,” *arXiv preprint arXiv:1812.10404*, 2018.
- [140] D. O’Connor, N. V. Potler, M. Kovacs, T. Xu, L. Ai, *et al.*, “The healthy brain network serial scanning initiative: A resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions,” *Gigascience*, vol. 6, no. 2, p. giw011, 2017.
- [141] 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.
- [142] 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.
- [143] 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.

- [144] 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.
- [145] 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.
- [146] 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.
- [147] L. H. Sobin, M. K. Gospodarowicz, and C. Wittekind, *TNM classification of malignant tumours*. John Wiley & Sons, 2011.
- [148] 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.
- [149] R. P. Mansour, “Visual charting method for creating electronic medical documents,” Apr. 16 2019. US Patent 10,262,106.
- [150] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, *et al.*, “Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 516–524, 2017.
- [151] 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.
- [152] 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.
- [153] 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.
- [154] G. Rossini, S. Parrini, T. Castroflorio, A. Deregibus, and C. L. Debernardi, “Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: A systematic review,” *Am. J. Orthod. Dentofacial Orthop.*, vol. 149, no. 2, pp. 161–170, 2016.
- [155] K. Huang, H. Cheng, Y. Zhang, *et al.*, “Medical knowledge constrained semantic breast ultrasound image segmentation,” in *Proc. International Conference on Pattern Recognition (ICPR)*, pp. 1193–1198, 2018.

- [156] C. Krittanawong, “The rise of artificial intelligence and the uncertain future for physicians,” *Eur. J. Intern. Med.*, vol. 48, pp. e13–e14, 2018.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] L. Rundo, C. Han, J. Zhang, R. Hataya, Y. Nagano, *et al.*, “CNN-based prostate zonal segmentation on T2-weighted MR images: A cross-dataset study,” in *Neural Approaches to Dynamics of Signal Exchanges*, Smart Innovation, Systems and Technologies, pp. 269–280, Springer, 2019.
- [161] E. H. P. Pooch, P. L. Ballester, and R. C. Barros, “Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification,” *arXiv preprint arXiv:1909.01940*, 2019.
- [162] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” *ACM TOG*, vol. 22, no. 3, pp. 313–318, 2003.
- [163] 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.
- [164] S. Chen, K. Ma, and Y. Zheng, “Med3D: Transfer learning for 3D medical image analysis,” *arXiv preprint arXiv:1904.00625*, 2019.
- [165] R. K. Samala, H. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha, “Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography,” *Med. Phys.*, vol. 43, no. 12, pp. 6654–6666, 2016.

Appendix A

Scientific Production

A.1 Related Publications/Presentations

Journal papers

- **C. Han**, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, Combining Noise-to-Image and Image-to-Image GANs: Brain MR Image Augmentation for Tumor Detection, *IEEE Access*, October 2019 (**Project 2**).
- **C. Han**, K. Murao, S. Satoh, H. Nakayama, Learning More with Less: GAN-based Medical Image Augmentation, *Medical Imaging Technology*, Japanese Society of Medical Imaging Technology, June 2019 (**Tutorial Paper**).

Book chapter

- **C. Han**, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, Infinite Brain MR Images: PGGAN-based Data Augmentation for Tumor Detection, In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, E. Pasero (eds.) *Neural Approaches to Dynamics of Signal Exchanges*, Springer, September 2019 (**Project 2**).

Conference proceedings

- **C. Han**, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, L. Rundo, H. Nakayama, S. Satoh, Learning More with Less: Conditional PGGAN-based Data Augmentation for Brain Metastases Detection Using Highly-Rough Annotation on MR Images, In *ACM International Conference on Information and Knowledge Management (CIKM)*, Beijing, China, November 2019 (**Project 3**).
- **C. Han**, Y. Kitamura, A. Kudo, A. Ichinose, L. Rundo, Y. Furukawa, K. Umemoto, H. Nakayama, Y. Li, Synthesizing Diverse Lung Nodules Wherever Massively: 3D Multi-Conditional GAN-based CT Image Augmentation for Object Detection, In *International Conference on 3D Vision (3DV)*, Québec City, Canada, September 2019 (**Project 4**).
- **C. Han**, H. Hayashi, L. Rundo, R. Araki, Y. Furukawa, W. Shimoda, S. Muramatsu, G. Mauri, H. Nakayama, GAN-based Synthetic Brain MR Image Generation, In *IEEE International Symposium on Biomedical Imaging (ISBI)*, Washington, D.C., The United States, April 2018 (**Project 1**).

Presentations

- **C. Han**, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, L. Rundo, H. Nakayama, S. Satoh, Learning More with Less: Conditional PGGAN-based MRI Augmentation with Highly-Rough Annotation for Brain Metastases Detection, 画像の認識・理解シンポジウム (*MIRU*), Osaka, Japan, July 2019.
- **C. Han** and H. Nakayama, Learning More with Less: Conditional PGGAN-based MRI Data Augmentation for Brain Tumor Detection, 東京大学生命科学シンポジウム (*BIO UT*), Tokyo, Japan, April 2019.
- **C. Han**, H. Hayashi, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, Towards Annotating Less Medical Images: PGGAN-based MR Image Augmentation for Brain Tumor Detection, 医用画像研究会 (*MI Symposium*), Okinawa, Japan, January 2019.

- **C. Han**, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, Infinite Brain Tumor Images: Can GAN-based Data Augmentation Improve Tumor Detection on MR Images?, 画像の認識・理解シンポジウム (*MIRU*), Hokkaido, Japan, August 2018.
- **C. Han**, H. Hayashi, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, Infinite Brain MR Images: PGGAN-based Data Augmentation for Tumor Detection, In *International Computer Vision Summer School (ICVSS)*, Sicily, Italy, July 2018.
- **C. Han**, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, Infinite Brain MR Images: PGGAN-based Data Augmentation for Tumor Detection, In *The Italian Workshop on Neural Networks (WIRN)*, Vietri sul Mare, Italy, June 2018.
- **C. Han**, H. Nakayama, Infinite Brain MR Images: PGGAN-based Data Augmentation for Tumor Detection, 東京大学生命科学シンポジウム (*BIO UT*), Tokyo, Japan, June 2018.
- 荒木 諒介, 韓 昌熙, 早志 英朗, Leonardo Rundo, MIRU2017 若手プログラム報告: GANによる合成脳MR画像生成, 情報処理学会 コンピュータビジョンとイメージメディア研究会 (*CVIM*), Osaka, Japan, May 2018.
- **C. Han**, R. Araki, W. Shimoda, S. Muramatsu, H. Hayashi, GAN-based Synthetic Medical Image Generation, 画像の認識・理解シンポジウム (*MIRU*) 若手プログラム, Hiroshima, Japan, August 2017.

A.2 Other Publications/Presentations

Journal paper

- L. Rundo*, **C. Han***, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi, M. C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama, P. Cazzaniga, USE-Net: incorporating Squeeze-and-Excitation

blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets, *Neurocomputing*, July 2019 (* denotes co-first authors).

Book chapters

- L. Rundo, **C. Han**, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M. S. Nobile, A. Tangherloni, M. C. Gilardi, S. Vitabile, H. Nakayama, G. Mauri, CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study, A. Esposito, M. Faundez-Zanuy, F. C. Morabito, E. Pasero (eds.) *Neural Approaches to Dynamics of Signal Exchanges*, Springer, September 2019.
- **C. Han**, K. Tsuge, H. Iba, Application of Learning Classifier Systems to Gene Expression Analysis in Synthetic Biology, In S. Patnaik, X. Yang, and K. Nakamatsu (eds.) *Nature Inspired Computing and Optimization: Theory and Applications*, Springer, March 2017.

Conference proceedings

- K. Murao, Y. Ninomiya, **C. Han**, K. Aida, S. Satoh, Cloud platform for deep learning-based CAD via collaboration between Japanese medical societies and institutes of informatics, In *SPIE Medical Imaging* (oral presentation), Houston, The United States, February 2020.
- **C. Han**, L. Rundo, K. Murao, Z. Á. Milacski, K. Umemoto, H. Nakayama, S. Satoh, GAN-based Multiple Adjacent Brain MRI Slice Reconstruction for Unsupervised Alzheimer's Disease Diagnosis, In *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, Bergamo, Italy, September 2019.
- **C. Han**, K. Tsuge, H. Iba, Optimization of Artificial Operon Construction by Consultation Algorithms Utilizing LCS, In *IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, Canada, July 2016.

Presentations

- **C. Han**, L. Rundo, J. Zhang, R. Hataya, Y. Nagano, G. Mauri, H. Nakayama, CNN-based Prostate Zonal Segmentation on Magnetic Resonance Images, 画像の認識・理解シンポジウム (*MIRU*), Hokkaido, Japan, August 2018.
- **C. Han**, F. Gesser, Z. Á. Milacski, L. Rundo, H. Nakayama, RNN-based Unsupervised Anomaly Detection for Alzheimer’s Disease Diagnosis Using Multiple MR Image Interpolation, In *International Summer School on Deep Learning (DeepLearn)*, Genoa, Italy, July 2018.
- L. Rundo, **C. Han**, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M. S. Nobile, A. Tangherloni, M. C. Gilardi, S. Vitabile, H. Nakayama, G. Mauri, CNN-based Prostate Zonal Segmentation on MR Images: A Multi-centric Study, In *International Summer School on Deep Learning (DeepLearn)*, Genoa, Italy, July 2018.
- L. Rundo, **C. Han**, J. Zhang, R. Hataya, Y. Nagano, C. Militello, C. Ferretti, M. S. Nobile, A. Tangherloni, M. C. Gilardi, S. Vitabile, H. Nakayama, G. Mauri, CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study, In *The Italian Workshop on Neural Networks (WIRN)*, Vietri sul Mare, Italy, June 2018.
- **C. Han**, J. Zhang, R. Hataya, Y. Nagano, H. Nakayama, L. Rundo, Prostate Zonal Segmentation Using Deep Learning, 医用画像研究会 (*MI Symposium*), Okinawa, Japan, March 2018.
- **C. Han**, F. Gesser, Z. Á. Milacski, CNN-based MRI Regression Using U-Net, 画像の認識・理解シンポジウム (*MIRU*), Hiroshima, Japan, August 2017.
- **C. Han**, MRI Reconstruction through CNN-based Regression for Brain Disease Diagnosis, 東京大学生命科学シンポジウム (*BIO UT*), Tokyo, Japan, April 2017.
- **C. Han**, F. Gesser, Z. Á. Milacski, Tomographic Slice Reconstruction by U-Net, In *Vision and Sports Summer School (VS3)*, Prague, Czech Republic, August

2016.

- **C. Han**, F. Navarro, F. Achilles, Tomographic Slice Reconstruction by Convolutional Neural Network, In *Medical Imaging Summer School (MISS)*, Sicily, Italy, August 2016.
- 韓昌熙, 柘植謙爾, 伊庭齊志, 機械学習による人工解糖系オペロン構築の最適化, 進化計算シンポジウム (*EC Symposium*), Aichi, Japan, December 2015.

A.3 Record of Awards

- International Conference on 3D Vision (3DV) Student Travel Grant (grant: \$CAD 500), September 2019.
- The University of Tokyo AI Solutions Global Competition 2019 Winner (prize: right to attend SU Global Summit 2020 including flight tickets and accommodation), August 2019.
- 画像の認識・理解シンポジウム (MIRU) 学生奨励賞 (Student Research Encouragement Award), August 2019.
- 電子情報通信学会 医用画像研究会 (MI Symposium) MI 研究奨励賞 (Best Research Award), May 2019.
- The University of Tokyo Graduate Program for Social ICT Global Creative Leaders (GCL) Mini-Presentation Competition Winner (prize: 0.1M Yen), December 2018.
- 画像の認識・理解シンポジウム (MIRU) 最優秀研究計画賞 (Best Research Plan Award, Team Leader, prize: GPUs worth 1M Yen), August 2017.
- Vision and Sports Summer School (VS3) Best Poster Award, August 2016.