Synthesizing Diverse Lung Nodules Wherever Massively:
3D Multi-Conditional GAN-based CT Image Augmentation for Object Detection

Changhee Han[1,3]   Yoshiro Kitamura[2]   Akira Kudo[2]   Akimichi Ichinose[2]   Leonardo Rundo[3]
Yujiro Furukawa[4]   Kazuki Umemoto[5]   Yuanzhong Li[2]   Hideki Nakayama[1]

[1]The University of Tokyo, Tokyo, Japan   [2]Fujifilm Corporation, Tokyo, Japan
[3]University of Cambridge, Cambridge, UK   [4]Jikei University School of Medicine, Tokyo, Japan
[5]Juntendo University School of Medicine, Tokyo, Japan

han@nlab.ci.i.u-tokyo.ac.jp

Abstract

Accurate Computer-Assisted Diagnosis, relying on large-scale annotated pathological images, can alleviate the risk of overlooking the diagnosis. Unfortunately, in medical imaging, most available datasets are small/fragmented. To tackle this, as a Data Augmentation (DA) method, 3D conditional Generative Adversarial Networks (GANs) can synthesize desired realistic/diverse 3D images as additional training data. However, 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. Therefore, we propose 3D Multi-Conditional GAN (MCGAN) to generate realistic/diverse $32 \times 32 \times 32$ nodules placed naturally on lung Computed Tomography images to boost sensitivity in 3D object detection. Our MCGAN adopts two discriminators 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/attenuation conditions. The results show that 3D Convolutional Neural Network-based detection can achieve higher sensitivity under any nodule size/attenuation at fixed False Positive rates and overcome the medical data paucity with the MCGAN-generated realistic nodules—even expert physicians fail to distinguish them from the real ones in Visual Turing Test.

1. Introduction

Accurate Computer-Assisted Diagnosis (CAD), thanks to recent Convolutional Neural Networks (CNNs), can al-
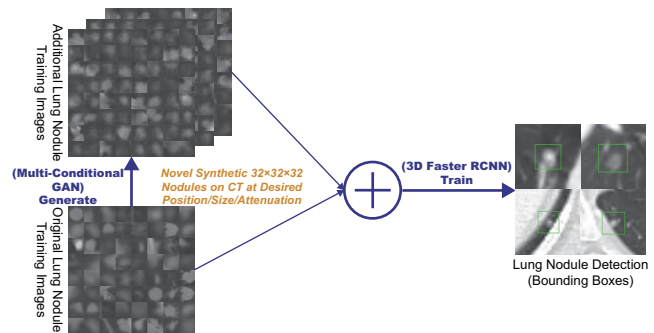

Figure 1: 3D MCGAN-based DA for better object detection: Our MCGAN generates realistic and diverse nodules naturally on lung CT scans at desired position/size/attenuation based on bounding boxes, and the CNN-based object detector uses them as additional training data.

leviate the risk of overlooking the diagnosis in a clinical environment. Such great success of CNNs, including diabetic eye disease diagnosis [12], 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 Data Augmentation (DA) techniques, such as geometric/intensity transformations [29, 23]. 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 its good generalization ability, Generative Adversarial Networks (GANs) [10] 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 [33].

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 (e.g., random noise samples to diverse pathological images) for classification [8, 16, 15]; others used image-to-image GANs (e.g., a benign image with a pathology-conditioning image to a malignant one) for object detection [14] and segmentation [3]. However, although 3D imaging is spreading in radiology (e.g., Computed Tomography (CT) and Magnetic Resonance Imaging), such 3D medical GAN-based DA approaches are limited, and mostly focus on segmentation [32, 18]—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.

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. 1), we propose 3D Multi-Conditional GAN (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 [34]. 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/attenuation conditions. We also evaluate the synthetic images' realism *via* Visual Turing Test [30] by two expert physicians, and visualize the data distribution *via* t-Distributed Stochastic Neighbor Embedding (t-SNE) [35]. The 3D MCGAN-generated additional training images can achieve higher sensitivity under any nodule size/attenuation at fixed False Positive (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 naturally 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.

2. Generative Adversarial Networks

GANs [10] have revolutionized image generation [20] *via* a two-player minimax game. However, difficult GAN training arises due to its two-player objective function, accompanying artifacts and mode collapse [11] when generating high-resolution images [27]—especially in 3D or conditional image generation; to tackle this, Wu *et al.* proposed 3D GAN [37] to generate realistic/diverse 3D objects *via* a mapping from a low-dimensional probabilistic space; Isola *et al.* proposed *Pix2Pix* GAN [17] to produce robust images using paired training samples; Park *et al.* proposed multi-conditional GAN [26] to generate 128×128 images from a base image and texts describing desired position. In this way, GANs can usually synthesize more realistic/diverse images than other common deep generative models, including variational autoencoders [19] suffering from the injected noise and imperfect reconstruction because of a single objective function [22]. Accordingly, as a DA method, most computer vision researchers chose GANs for improving classification [1], object detection [25], and segmentation [39] to overcome the training data paucity.

Also in medical imaging, to facilitate object detection and segmentation, researchers usually used conditional GANs to generate medical images at desired positions for DA. Han *et al.* generated 256×256 brain MR images with tumors at desired positions/sizes for tumor detection [14]. As 3D GANs for DA, Jin *et al.* [18] generated $64 \times 64 \times 64$ CT images of both nodules and surrounding tissues—unlike we only generate $32 \times 32 \times 32$ nodules located smoothly on surroundings—for 2D nodule segmentation. Gao *et al.* [9] generated $40 \times 40 \times 18$ 3D subvolumes of nodules for subvolume-based 3D nodule detection *via* binary classification; but, the subvolume-based detection accompanies numerous FPs, and unlike our work, most other 3D object de-
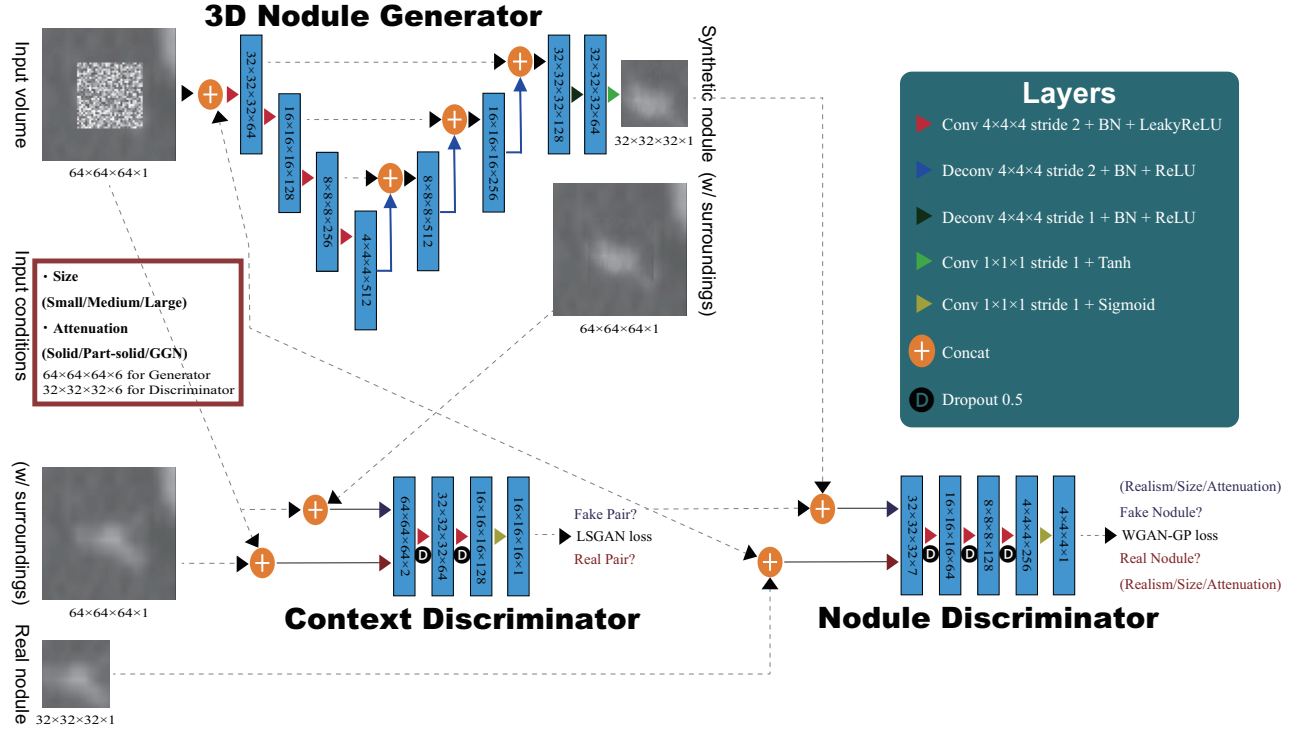
Figure 2: 3D MCGAN architecture for realistic and diverse $32 \times 32 \times 32$ lung nodule generation: The context discriminator learns to classify real *vs* synthetic nodule/surrounding pairs while the nodule discriminator learns to classify real *vs* synthetic nodules.

tectors cannot use the generated nodules as additional training data since they do not condition nodule positions.

To the best of our knowledge, our work is the first 3D medical GAN-based DA approach using automatic bounding box annotation while 3D bounding boxes require much cheaper annotation cost than rigorous 3D segmentation. Moreover, we, for the first time, generate 3D multi-conditional images using GANs. In terms of annotation cost, generating realistic and diverse $32 \times 32 \times 32$ lung nodules at desired position/size/attenuation using 3D MCGANs—may become a clinical breakthrough .

3. Methods

3.1. 3D MCGAN-based Image Generation

Data Preparation This study exploits the Lung Image Database Consortium image collection (LIDC) dataset [2] containing $1,018$ chest CT scans with lung nodules. Since the American College of Radiology recommends lung nodule evaluation using thin-slice CT scans [31], we only use scans with the slice thickness ≤ 3 mm and 0.5 mm $\leq$ in-plane pixel spacing ≤ 0.9 mm. Then, we interpolate the slice thickness to 1.0 mm and exclude scans with slice number > 400.

To explicitly provide MCGAN with meaningful nodule appearance information and thus boost DA performance, the authors further annotate those nodules by size and attenuation for GAN training with multiple conditions: small

(slice thickness ≤ 10 mm); medium (10 mm $\leq$ slice thickness ≤ 20 mm); large (slice thickness > 20 mm); solid; part-solid; Ground-Glass Nodule (GGN). Afterwards, the remaining dataset (745 scans) is divided into: (*i*) a training set (632 scans/$3,727$ nodules); (*ii*) a validation set (37 scans/143 nodules); (*iii*) a test set (76 scans/265 nodules); only the training set is used for MCGAN training to be methodologically sound. 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. 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 Volumes of Interest (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 [5]-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 [17]-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 (LSGAN) [21]; the nodule discriminator attempts to classify real *vs* synthetic nodules with size/attenuation conditions using Wasserstein loss with Gradient Penalty (WGAN-GP) [11]. The LSGAN 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 [25] 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 [9] to generate realistic images, even modifying it for further realism [18]. 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:

$$\begin{aligned} G^* &= \arg\min_G \max_{D1,D2} \mathcal{L}_{\mathbf{LSGAN}}(G, D1) \\ &\quad + \mathcal{L}_{\mathbf{WGAN\text{-}GP}}(G, D2), \end{aligned} \tag{1}$$

$$\begin{aligned} G^* &= \arg\min_G \max_{D1,D2} \mathcal{L}_{\mathbf{LSGAN}}(G, D1) \\ &\quad + \mathcal{L}_{\mathbf{WGAN\text{-}GP}}(G, D2) + 100\mathcal{L}_{\ell_1}(G). \end{aligned} \tag{2}$$

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 [17, 25].
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.

3.2. Lung Nodule Detection Using 3D Faster RCNN

3D Faster RCNN is a 3D version of Faster RCNN [28] using multi-task loss with a 27-layer Region Proposal Network of 3D convolutional layers, batch normalization layers, and 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 [24], 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 $(1.0 \times 10^{-4}$ 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 Intersection over Union (IoU) threshold 0.25/detection threshold 0.5 to avoid severe FPs.

3.3. Clinical Validation Using Visual Turing Test

To quantitatively evaluate the realism of MCGAN-generated images, we supply, in a 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. Such Visual Turing Test [30] can evaluate the visual quality of GAN-generated medical images in a clinical environment, where physicians' specialty is critical [13, 4].

3.4. Visualization Using t-SNE

To visually analyze the distribution of real/synthetic images, we use t-SNE [35] on a random selection of 500 real, 500 synthetic, and 500 ℓ_1 loss-added synthetic nodule im-

Lung CT (Real nodule w/ surroundings, 64 × 64 × 64)

Lung CT (Noise box-replaced nodule w/ surroundings, 64 × 64 × 64)

Lung CT (Synthetic nodule w/ surroundings, 64 × 64 × 64)

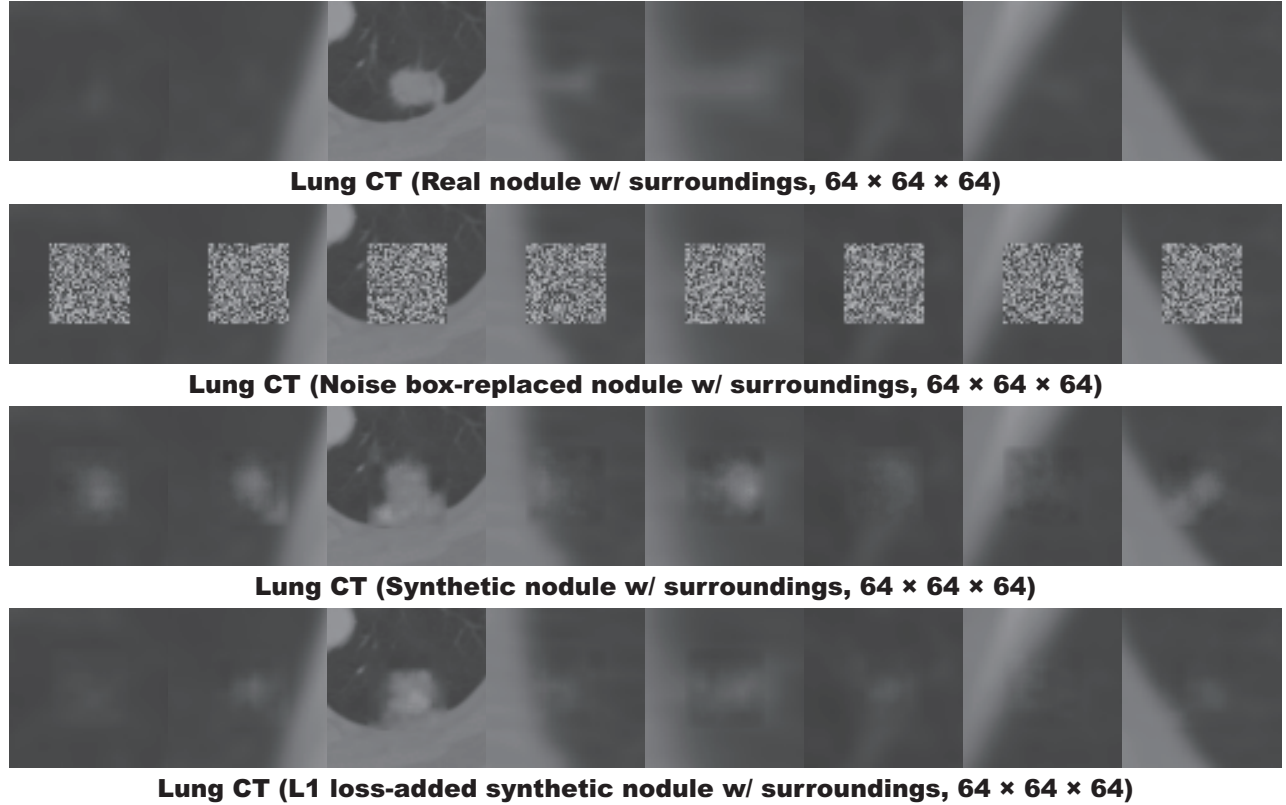**Lung CT (L1 loss-added synthetic nodule w/ surroundings, 64 × 64 × 64)**

Figure 3: 2D axial view of example $64 \times 64 \times 64$ lung nodules with surrounding tissues; 3D MCGANs generate only $32 \times 32 \times 32$ nodules.

Table 1: Nodule detection results (CPM) with/without 3D MCGAN-based DA for 3D Faster RCNN under IoU ≥ 0.25. Both results without/with ℓ_1 loss at different augmentation ratio are compared. CPM is average sensitivity at $1/8, 1/4, 1/2, 1, 2, 4$, and 8 FPs per scan.

	CPM	CPM by Size			CPM by Attenuation		
		Small	Medium	Large	Solid	Part-solid	GGN
632 real images	0.518	0.447	0.618	0.624	0.655	0.464	0.242
+ $1\times$ 3D MCGAN-based DA	**0.550**	**0.452**	**0.683**	**0.662**	**0.699**	0.521	0.244
+ $2\times$ 3D MCGAN-based DA	0.527	0.447	0.674	0.429	0.655	0.407	**0.289**
+ $3\times$ 3D MCGAN-based DA	0.512	0.411	0.644	0.662	0.616	**0.579**	0.277
+ $1\times$ 3D MCGAN-based DA w/ ℓ_1	0.508	0.430	0.633	0.556	0.626	0.471	0.271
+ $2\times$ 3D MCGAN-based DA w/ ℓ_1	0.509	0.406	0.644	0.654	0.649	0.436	0.233
+ $3\times$ 3D MCGAN-based DA w/ ℓ_1	0.479	0.389	0.594	0.617	0.596	0.507	0.226

ages, with a perplexity of 100 for $1,000$ iterations to get a 2D representation. We normalize the input images to $[0, 1]$. The t-SNE method represents high-dimensional data into a lower-dimensional space by reducing the dimensionality; it uses perplexity to non-linearly balance between the input data's local and global aspects.

4. Results

4.1. Lung Nodules Generated by 3D MCGAN

We generate realistic nodules in noise box regions at various position/size/attenuation, naturally blending with sur-

rounding tissues including vessels, soft tissues, and thoracic walls (Fig. 3). Especially, when trained without ℓ_1 loss, those synthetic nodules look much more different from the original real ones, including slight shading difference.

4.2. Lung Nodule Detection Results

Table 1 and Fig. 4 show that it is easier to detect nodules with larger size and 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 under any
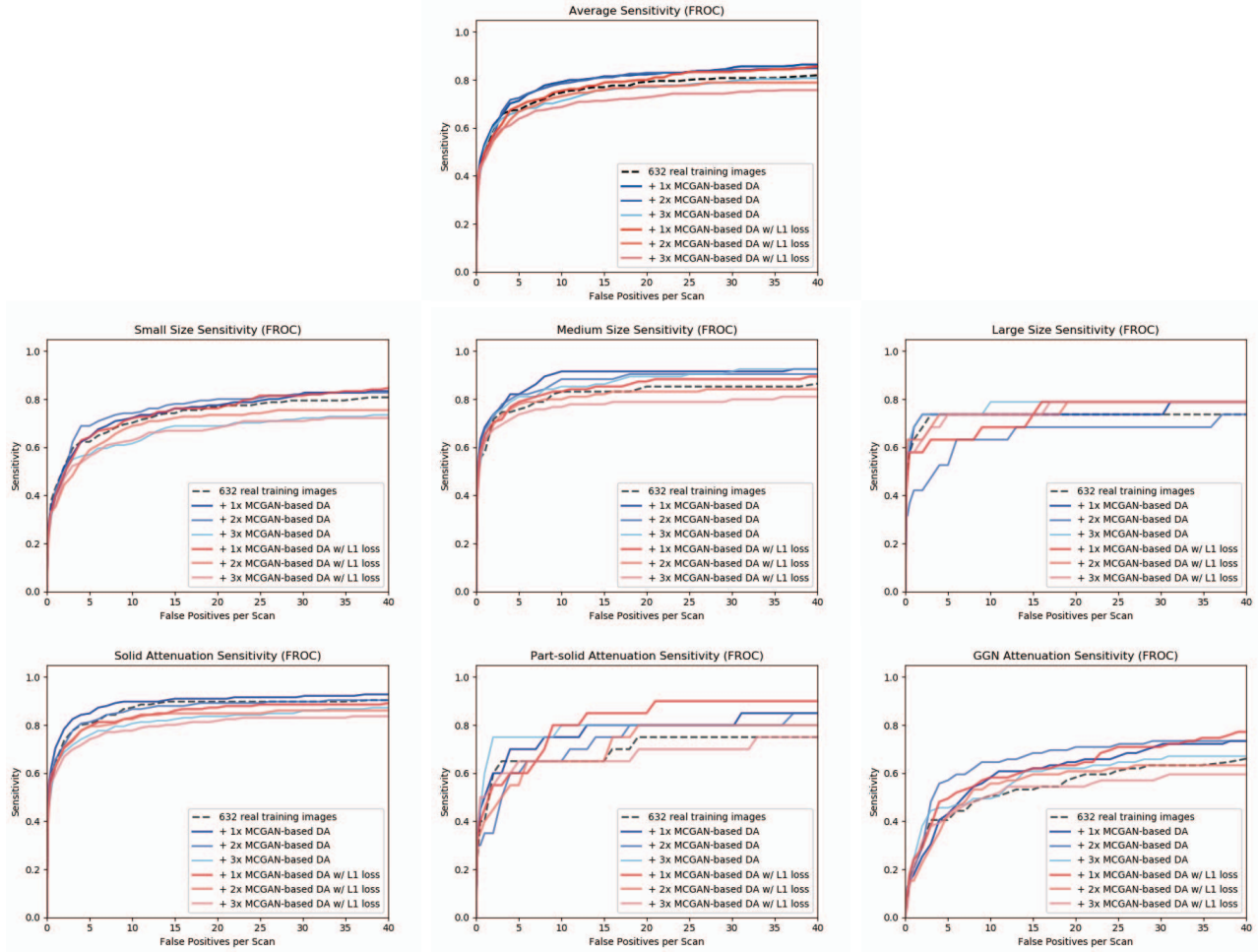
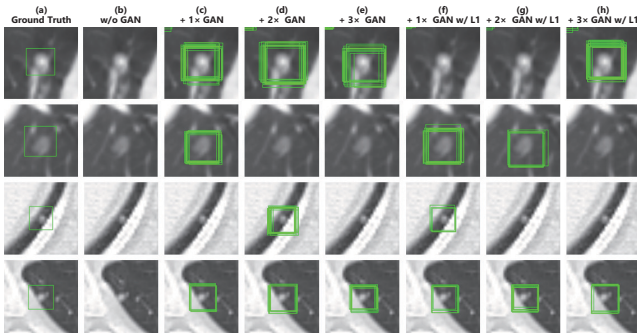Figure 4: FROC curves of different DA setups by average/size/attenuation.

Figure 5: Example detection results with 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.

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. 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, probably 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.

4.3. Visual Turing Test Results

As Table 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' re-

Table 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.

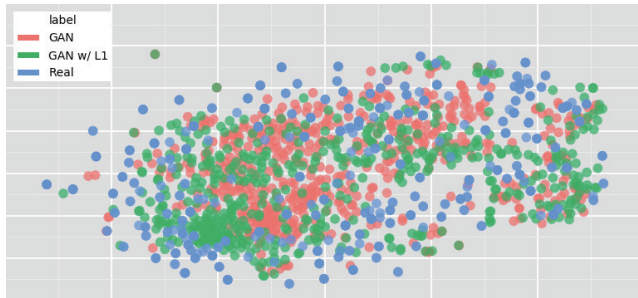|		Accuracy	Real Selected as Real	Real as Synt	Synt as Real	Synt as Synt
Test1	Physician1	43%	19	31	26	24
	Physician2	43%	13	37	20	30
Test2	Physician1	57%	22	28	15	35
	Physician2	53%	11	39	8	42
Test3	Physician1	62%	25	25	13	37
	Physician2	79%	32	18	3	47
Test4	Physician1	58%	21	29	13	37
	Physician2	66%	36	14	20	30

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

alism, 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 [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. 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.

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 un-

der 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 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 [38] and cutout [6]. 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 [36]. 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 [7].

References

[1] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[2] 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. *Medical physics*, 38(2):915–931, 2011.

[3] O. Bailo, D. Ham, and Y. M. Shin. Red blood cell image generation for data augmentation using conditional generative adversarial networks. *arXiv preprint arXiv:1901.06219*, 2019.

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

[5] 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)*, pages 424–432. Springer, 2016.

[6] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[7] 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:1812.01547*, 2018.

[8] 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*, 321:321–331, 2018.

[9] C. Gao, S. Clark, J. Furst, and D. Raicu. Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 109501K, 2019.

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

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

[12] 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*, 316(22):2402–2410, 2016.

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

[14] 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)*, 2019. (In press).

[15] C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi. Infinite brain MR images: PGGAN-based data augmentation for tumor detection. In *Neural Approaches to Dynamics of Signal Exchanges*, volume 151 of *Smart Innovation, Systems and Technologies*. Springer, 2018. (In press).

[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. *arXiv preprint arXiv:1905.13456*, 2019.

[17] 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)*, pages 1125–1134, 2017.

[18] 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)*, pages 732–740, 2018.

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc. International Conference on Learning Representations (ICLR) arXiv:1312.6114*, 2013.

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

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

[22] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: unifying variational autoencoders and generative adversarial networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 2391–2400, 2017.

[23] F. Milletari, N. Navab, and S.-A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. IEEE International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[24] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken. On combining computer-aided detection systems. *IEEE Transactions on Medical Imaging*, 30(2):215–223, 2011.

[25] 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.

[26] 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:1805.01123*, 2018.

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adver-

sarial networks. In *Proc. International Conference on Learning Representations (ICLR) arXiv:1511.06434*, 2016.

[28] 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)*, pages 91–99, 2015.

[29] 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)*, pages 234–241, 2015.

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

[31] A. A. A. Setio, A. Traverso, T. De Bel, M. S. 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. *Medical image analysis*, 42:1–13, 2017.

[32] 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)*, pages 1–11, 2018.

[33] 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)*, pages 2107–2116, 2017.

[34] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

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

[36] 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:1903.03496*, 2019.

[37] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016.

[38] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[39] Y. Zhu, M. Aoun, M. Krijn, J. Vanschoren, and H. T. Campus. Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants. In *Proc. Computer Vision Problems in Plant Phenotyping (CVPPP)*, 2018.