2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

GAN-BASED SYNTHETIC BRAIN MR IMAGE GENERATION

Changhee Han[1], *Hideaki Hayashi*[2], *Leonardo Rundo*[3], *Ryosuke Araki*[4], *Wataru Shimoda*[5]
Shinichi Muramatsu[6], *Yujiro Furukawa*[7], *Giancarlo Mauri*[3], *Hideki Nakayama*[1]

[1]Grad. School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2]Dept. of Advanced Information Technology, Kyushu University, Fukuoka, Japan
[3]Dept. of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[4]Grad. School of Engineering, Chubu University, Aichi, Japan
[5]Dept. of Informatics, The University of Electro-Communications, Tokyo, Japan
[6]Grad. School of Science and Technology, Shinshu University, Nagano, Japan
[7]Kanto Rosai Hospital, Kanagawa, Japan

ABSTRACT

In medical imaging, it remains a challenging and valuable goal how to generate realistic medical images completely different from the original ones; the obtained synthetic images would improve diagnostic reliability, allowing for data augmentation in computer-assisted diagnosis as well as physician training. In this paper, we focus on generating synthetic multi-sequence brain Magnetic Resonance (MR) images using Generative Adversarial Networks (GANs). This involves difficulties mainly due to low contrast MR images, strong consistency in brain anatomy, and intra-sequence variability. Our novel realistic medical image generation approach shows that GANs can generate 128×128 brain MR images avoiding artifacts. In our preliminary validation, even an expert physician was unable to accurately distinguish the synthetic images from the real samples in the Visual Turing Test.

Index Terms— Generative Adversarial Networks, Synthetic Medical Image Generation, Brain MRI, Data Augmentation, Physician Training, Visual Turing Test

1. INTRODUCTION

Along with classic methods [1], Convolutional Neural Networks (CNNs) have recently revolutionized medical image analysis [2], including brain Magnetic Resonance Imaging (MRI) segmentation [3]. However, CNN training demands extensive medical data that are laborious to obtain [4]. To overcome this issue, data augmentation techniques via reconstructing original images are common for better performance, such as geometry and intensity transformations [5, 6].

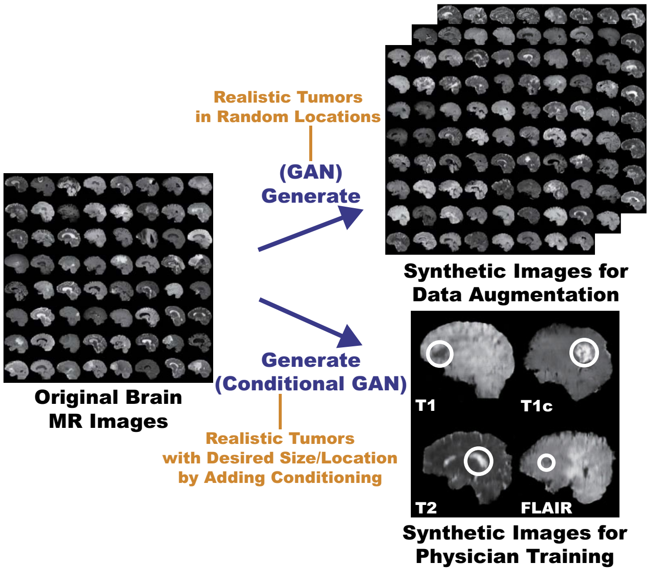

Fig. 1. Potential applications of the proposed GAN-based synthetic brain MR image generation: (1) data augmentation for better diagnostic accuracy by generating random realistic images giving insights in classification; (2) physician training for better understanding various diseases to prevent misdiagnosis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

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, Generative Adversarial Network (GAN)-based data augmentation has excellently performed in general computer vision tasks. It attributes to GAN's good generalization ability from matching the generated distribution from noise variables 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 [7].

So, how can we generate realistic medical images completely different from the original samples? Our aim is to gen-

This work was partially supported by the Graduate Program for Social ICT Global Creative Leaders of The University of Tokyo by JSPS.

978-1-5386-3636-7/18/$31.00 ©2018 IEEE

erate synthetic multi-sequence brain MR images using GANs, which is essential in medical imaging to increase diagnostic reliability, such as via data augmentation in computer-assisted diagnosis as well as physician training and teaching (Fig. 1) [8]. 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 data augmentation adopts Deep Convolutional GAN (DCGAN) [9] and Wasserstein GAN (WGAN) [10] to generate realistic images, and an expert physician validates them via the Visual Turing Test [11].

Research Questions. We mainly address two questions:

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

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: data augmentation and physician training.
- **Medical Image Generation:** This research provides how to exploit medical images with intrinsic intra-sequence variability towards GAN-based data augmentation for medical imaging.

2. GENERATIVE ADVERSARIAL NETWORKS

Since the breakthrough paper by Goodfellow *et al.* in 2014 [12], GANs have shown promising results for image generation in general computer vision [13]. GANs generate highly realistic images, without a well-defined objective function associated with difficult training accompanying oscillations and mode collapse—i.e., a common failure case where the generator learns with extremely low variety. Whereas Variational Autoencoders (VAEs) [14], the other most used deep generative models, have an objective likelihood function to optimize, and could so generate blurred samples because of the injected noise and imperfect reconstruction [15].

Therefore, many medical imaging researchers have begun to use GANs recently, such as in image super-resolution [16], anomaly detection [17], and estimating CT images from the corresponding MR images [18]. As GANs allow adding conditioning on the class labels and images, they often use such conditional GANs to produce desired images, while it makes learning robust latent spaces difficult.

Differently from a very recent work of GANs for biological image synthesis (fluorescence microscopy) [19], to the best of our knowledge, this is the first GAN-based realistic brain tumor MR image generation approach aimed at data augmentation and physician training. Instead of reconstructing real brain MR images themselves with respect to geometry/intensity, a completely different approach—generating novel realistic images using GANs—may become a clinical breakthrough.

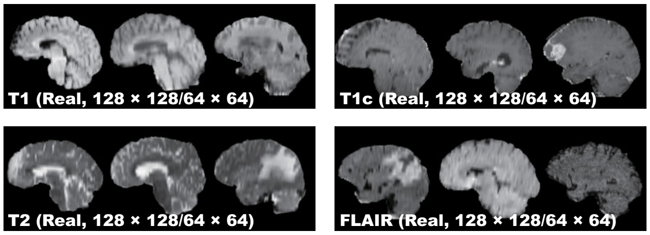
T1 (Real, 128 × 128/64 × 64)
T1c (Real, 128 × 128/64 × 64)
T2 (Real, 128 × 128/64 × 64)
FLAIR (Real, 128 × 128/64 × 64)

Fig. 2. Example real MR images used for training the GANs: the resized sagittal multi-sequence brain MRI scans of patients with HGG on the BRATS 2016 training dataset [20].

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.

3.1. The BRATS 2016 Dataset

This paper 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 [20]. 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 (T2), and Fluid Attenuation Inversion Recovery (FLAIR) sequences—they were skull stripped and resampled to isotropic $1\text{mm} \times 1\text{mm} \times 1\text{mm}$ resolution with image dimension $240 \times 240 \times 155$; among the different sectional planes, we use the sagittal multi-sequence scans of patients with HGG 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.

3.2. Proposed GAN-based Image Generation Approach

3.2.1. 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 64×64 and 128×128 from 240×155 for better GAN training (DCGAN architecture results in stable training on 64×64 [9], and so 128×128 is reasonably a high-resolution). Fig. 2 shows some real MR images used for training; each sequence contains 15,400 images with 220 patients × 70 slices (61,600 in total).

3.2.2. *GAN-based 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 data augmentation, containing the features of all four sequences. We also generate 64×64 Concat images to compare the generation performance in terms of image size.

DCGAN. DCGAN [9] is a standard GAN [12] with a convolutional architecture for unsupervised learning; this generative model uses up-convolutions interleaved with ReLu non-linearity and batch-normalization.

Let p_{data} be a generating distribution over data $\boldsymbol{x}$. The generator $G(\boldsymbol{z}; \theta_g)$ is a mapping to data space that takes a prior on input noise variables $p_{\boldsymbol{z}}(\boldsymbol{z})$, where G is a neural network with parameters θ_g. Similarly, the discriminator $D(\boldsymbol{x}; \theta_d)$ is a neural network with parameters θ_d that takes either real data or synthetic data and outputs a single scalar probability that $\boldsymbol{x}$ 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 $V(G, D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\ + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (1)$$

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 $p_{\boldsymbol{z}}$ and G.

DCGAN Implementation Details. We use the same DCGAN architecture [9] 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. WGAN [10] 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 (a.k.a. the Wasserstein-1 metrics):

$$W(p_g, p_r) = \inf_{p \in \prod(p_g, p_r)} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{x}') \sim p} \|\boldsymbol{x} - \boldsymbol{x}'\|, \quad (2)$$

where $\prod(p_g, p_r)$ 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.

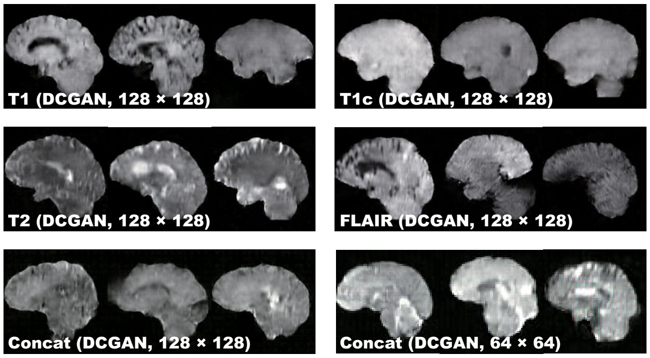

Fig. 3. Example synthetic MR images yielded by DCGAN.

WGAN Implementation Details. We use the same DCGAN architecture [9] for WGAN training. A batch size of 64 and Root Mean Square Propagation (RMSprop) optimizer with 5.0×10^{-5} learning rate were implemented.

3.3. Clinical Validation Using the 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 a 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 [11] 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 [7]. This applies also to medical images in clinical environments [21], wherein physicians' expertise is critical.

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. The training took about 2 (1) hours to train each 128×128 (64×64) sequence on an Nvidia GeForce GTX 980 GPU, increasingly learning realistic features.

4.1. MR Images Generated by GANs

DCGAN. Fig. 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.

Table 1. Visual Turing Test results by a physician for classifying real vs synthetic images. It should be noted that proximity to 50% of accuracy indicates superior performance (chance = 50%).

	Accuracy (%)	**Real Selected as Real**	**Real as Synt**	**Synt as Real**	**Synt as Synt**
T1 (DCGAN, 128 × 128)	70	26	24	6	44
T1c (DCGAN, 128 × 128)	71	24	26	3	47
T2 (DCGAN, 128 × 128)	64	22	28	8	42
FLAIR (DCGAN, 128 × 128)	54	12	38	8	42
Concat (DCGAN, 128 × 128)	77	34	16	7	43
Concat (DCGAN, 64 × 64)	54	13	37	9	41
T1 (WGAN, 128 × 128)	64	20	30	6	44
T1c (WGAN, 128 × 128)	55	13	37	8	42
T2 (WGAN, 128 × 128)	58	19	31	11	39
FLAIR (WGAN, 128 × 128)	62	16	34	4	46
Concat (WGAN, 128 × 128)	66	31	19	15	35
Concat (WGAN, 64 × 64)	53	18	32	15	35

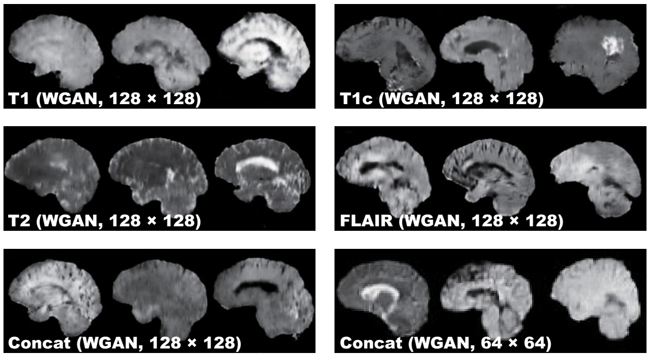

Fig. 4. Example synthetic MR images yielded by WGAN.

WGAN. Fig. 4 shows the example output of WGAN in each sequence. Outperforming remarkably DCGAN, WGAN successfully captures the sequence-specific texture and the appearance of the tumors 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 the boundaries of the brain, due to the introduction of unexpected intensity patterns.

4.2. Visual Turing Test Results

Table 1 shows the confusion matrix concerning the Visual Turing Test. Even the expert physician found classifying real and synthetic images challenging, especially in lower resolution due to their less detailed appearances 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%).

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 data augmentation 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 the inferior realism and mode collapse in terms of intensity. We only use the slices of interest in training to obtain desired MR images and generate both original/Concat sequence images for data augmentation 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) [22], which assesses whether two samples are drawn from the same distribution. Currently this work uses sagittal MR images alone, so we will generate coronal and transverse images in the near future. 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 data augmentation, while realistic images give more insights on geometry/intensity transformations in classification, more realistic images do not always assure better data augmentation, 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 extensively the latent spaces of GANs.

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.

6. REFERENCES

[1] L. Rundo, C. Militello, G. Russo, et al., “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.

[2] D. Shen, G. Wu, and H.I. Suk, “Deep learning in medical image analysis,” *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

[3] M. Havaei, A. Davy, D. Warde-Farley, et al., “Brain tumor segmentation with deep neural networks,” *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[4] D. Ravì, C. Wong, F. Deligianni, et al., “Deep learning for health informatics,” *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, 2017.

[5] 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)*. Springer, 2015, pp. 234–241.

[6] 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)*, 2016, pp. 565–571.

[7] A. Shrivastava, T. Pfister, O. Tuzel, et al., “Learning from simulated and unsupervised images through adversarial training,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2107–2116.

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

[9] 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.

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

[11] 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)*, 2016, pp. 2234–2242.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., “Generative adversarial nets,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[13] J. Zhu, T. Park, P. Isola, et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE International Conference on Computer Vision (ICCV), arXiv preprint arXiv:1703.10593*, 2017.

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

[15] L. Mescheder, S. Nowozin, and A. Geiger, “Adversarial variational Bayes: unifying variational autoencoders and generative adversarial networks,” *arXiv preprint arXiv:1701.04722*, 2017.

[16] D. Mahapatra, B. Bozorgtabar, S. Hewavitharanage, et al., “Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2017, pp. 382–390.

[17] T. Schlegl, P. Seeböck, S.M. Waldstein, et al., “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *Proc. International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[18] D. Nie, R. Trullo, J. Lian, et al., “Medical image synthesis with context-aware generative adversarial networks,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2017, pp. 417–425.

[19] A. Osokin, A. Chessel, R.E. Carazo Salas, et al., “GANs for biological image synthesis,” in *Proc. International Conference on Computer Vision (ICCV), arXiv preprint arXiv:1708.04692v2*, 2017.

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

[21] M. J. M. Chuquicusma, S. Hussein, J. Burt, et al., “How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis,” *arXiv preprint arXiv:1710.09762v1*, 2017.

[22] 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.