Gun Ahn ORCID iD: 0000-0002-4185-6564

Byung Sun Choi ORCID iD: 0000-0002-4492-4358

Hyuk-Soo Han ORCID iD: 0000-0003-1229-8863

Du Hyun Ro ORCID iD: 0000-0001-6199-908X

High-Resolution Knee Plain Radiography Image Synthesis Using Style Generative Adversarial Network Adaptive Discriminator Augmentation

Gun Ahn[1,2*], Byung Sun Choi[2*], Sunho Ko[3], Changwung Jo[3], Hyuk-Soo Han[2], Myung Chul Lee[2], Du Hyun Ro[2]

1. Interdisciplinary Program of Bioengineering, Seoul National University

2. Department of Orthopedic Surgery, Seoul National University Hospital

3. Department of Medicine, Seoul National University

(*) denotes equal contribution

Correspondence to:

Du Hyun Ro, MD

101 Daehak-ro, Ihwa-dong, Jongno-Gu, Seoul, 03080, Contact: +82-10-6312-4610, Email.

duhyunro@gmail.com

Running Title

Knee Image Synthesis Using GAN

Author contribution

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jor.25325.

Gun Ahn and Du Huyn Ro designed the study. Gun Ahn performed a study. Gun Ahn, Byung Sun Choi, Du Hyun Ro performed the analysis and interpretation of the results. Sunho Ko, Changwung Jo, Hyuk-Soo Han, and Myung Chul Lee worked on data collection. Each author has contributed to the writing and revising of the manuscript. All authors have read and approved the final submitted manuscript

Abstract

In this retrospective study, 10,000 anteroposterior (AP) radiography of the knee from a single institution was used to create medical dataset that are more balanced and cheaper to create. Two types of convolutional networks were used, deep convolutional GAN (DCGAN) and Style GAN Adaptive Discriminator Augmentation (StyleGAN2-ADA). To verify the quality of generated images from StyleGAN2-ADA compared to real ones, the Visual Turing test was conducted by two computer vision experts, two orthopedic surgeons, and a musculoskeletal radiologist. For quantitative analysis, the Fréchet inception distance (FID), and principal component analysis (PCA) were used. Generated images reproduced the features of osteophytes, joint space narrowing, and sclerosis. Classification accuracy of the experts was 34, 43, 44, 57, and 50%. FID between the generated images and real ones was 2.96, which is significantly smaller than another medical dataset (BreCaHAD = 15.1). PCA showed that no significant difference existed between the PCs of the real and generated images ($P > 0.05$). At least 2000 images were required to make reliable images optimally. By performing PCA in latent space, we were able to control the desired PC that show a progression of arthritis. Using a GAN, we were able to generate knee X-ray images that accurately reflected the characteristics of the arthritis progression stage, which neither human experts, nor artificial intelligence could discern apart from

the real images. In summary, our research opens up the potential to adopt a generative model to synthesize realistic anonymous images that can also solve data scarcity and class inequalities.

Level of evidence: III

Introduction

The combination of medical big data and artificial intelligence is expected to bring significant innovations to healthcare as a predictable technology with high accuracy that extends beyond analysis.[1] High-quality big data plays a critical role in developing models that can be used in clinical settings.

However, there are obvious limitations in the collection and utilization of medical big data. Since each hospital lacks consistency in the form, storage method, and terminology used. In addition, due to the nature of diseases, the problem of data imbalance also arises.[2] Most importantly, data accessibility is extremely low. Inevitably, the usefulness of the data and the privacy risks are related to each other's trade-offs.[3,4] Medical researchers need to carry on numerous paperwork and sometimes they have to visit a data center physically to access necessary data.

Generative adversarial networks (GANs) have shown outstanding results for data augmentation. In the computer vision community, GANs have generated car images,[5] faces,[6] and animal faces. Also, synthetic medical data using GANs have been used in various studies in the medical field. Since deep learning demonstrated the potential to complement image interpretation and augment image representation and classification, deep learning such as GAN is widely adopted in medical imaging research.[7] For instance, Woltering et al. improved computed tomography (CT) image quality with help of GANs.[8] Frid-Adar et al. used deep convolutional generative adversarial

networks (DCGAN) to synthesize images for three classes of liver lesions including cysts, metastases, and hemangiomas.[9] Other researchers used progressive GAN (PGAN) to synthesize realistic high-resolution skin lesion images that dermatologists cannot classify with real images.[8,10] Also, Bermudez et al. generated brain magnetic resonance imaging (MRI) images to be of comparable quality to real ones, which even neuroradiologists found to be of comparable quality to the real images.[11]

Osteoarthritis (OA) of the knee is the most common joint disorder[12] and cause of disability among adults.[13] Plain radiographs of the knee are a key aspect of grading OA joints.[14] The radiographic signs show evidence needed to determine the grading of OA severity in joints such as joint space narrowing, osteophytes, subchondral bone sclerosis, and joint subluxation.[15,16] The Kellgren and Lawrence (KL) system, using five grades concerning the above features, is the common method for classifying OA severity by simple knee anteroposterior (AP) X-ray.[17] The KL scoring is related to the progression of OA, and it is considered to be a great help to determine the appropriate clinical treatment with symptoms and clinical functions.[18] Bowes et al. points out that KL grade is associated with risk of current and future pain, functional limitation, and total knee replacement.[19] Carlson et al. highlight the needs of following recommended surgery upon KL grade. Total knee arthroplasty (TKA) is recommended in KL grade 4, and the recommended treatments are different in each stage in the American Academy of Orthopaedic Surgeons (AAOS) guideline. As revealed in previous research, radiographic severity cannot be considered as an independent factor in determining the surgical treatment, but should be considered together with symptoms and clinical functions.[20]

In this research, we aimed to generate realistic knee X-ray images with the help of a GAN. Then, we checked whether human experts[17] or algorithms were able to

distinguish between the real and generated images. OA severity is compared to features used in the KL scoring system in generated images, so later we could generate a dataset with KL grade. Lastly, we wanted to determine under what conditions reliable images could be generated. With the model, researchers could use medical images in industries and research without worrying about privacy issues.

Materials and Methods

Dataset

This study was approved by our institutional review board (IRB: 2009-181-1161). This retrospective study included 10,000 consecutive patients who got both knees standing AP radiographs at our institution (Seoul National University Hospital, SNUH). Patients who had previous bone surgery on lower extremities such as arthroplasty, osteotomy, and fracture surgery were excluded. Both knees standing AP were obtained with the knees fully extended and patellae facing forward. The beam was projected 6-10 degrees caudally depending on the patients. The images were center-cropped to square size and down-sampled to 512×512-pixel. No further preprocessing was applied.

Generative Adversarial Networks (GANs)

The original GAN framework was introduced for estimating generative models via an adversarial process, in which two models are simultaneously trained, a generator that captures the data distribution, and a discriminator that estimates the probability that a sample came from the training data rather than from the generator. Ultimately, the goal is to teach the generator to approximate the training dataset distribution. The training procedure for the generator maximizes the probability of the discriminator

making a mistake while the discriminator tries to minimize the mistakes. Therefore, this framework corresponds to a mini-max two-player game.[21] In this work, we employed two different GAN concepts for the task of knee X-ray image synthesis, DCGAN, and style generative adversarial networks adaptive discriminator augmentation (StyleGAN2-ADA).

Deep Convolutional Generative Adversarial Networks (DCGAN)

The DCGAN architecture is one of the most popular convolutional GANs. DCGAN is a network that trains discriminators for image classification tasks, showing competitive performance with other unsupervised algorithms. The architecture is designed with concepts like specific weight initialization, allowing robust training and leaky-ReLu activations to avoid sparse gradients.

Style Generative Adversarial Networks Adaptive Discriminator Augmentation (StyleGAN2-ADA)

StyleGAN2-ADA is well known for showing the best performance in image generation tasks for the Flickr-Faces-HQ(FFHQ)[6] dataset according to the paperwithcode website. The StyleGAN2-ADA adopted an adaptive discriminator augmentation (ADA) mechanism that significantly stabilizes training in limited data regimes, which was perfectly suitable for our dataset. The ADA mechanism does not require changes to loss functions or network architectures and is applicable both when training from scratch and when fine-tuning an existing GAN on another dataset. The researchers demonstrated on several datasets that good results were possible using only a few thousand training images, often matching the best state-of-the-art network results with an order of magnitude fewer images.[22]

Image Synthesis

We trained a StyleGAN2-ADA from all 10,000 images, as well as a DCGAN. The DCGAN was trained for 500 epochs. Hyperparameters are as follows. Batch size during training was 128, size of z latent vector was 100, the base size of feature maps in the generator was 16, the base size of feature maps in discriminator was 16, learning rate for optimizers was 0.0002, and beta1 hyperparameter for Adam optimizers was 0.5. Style GAN2-ADA was trained until 1M augmented images were reached and we used default settings that populated dynamically based on resolution and GPU count. In this study, DCGAN image generation takes around 5 hours, whereas StyleGAN2-ADA takes 20 hours with 2 NVIDIA GeForce RTX 3090 GPUs. All code for generating gray images with DCGAN and StyleGAN2-ADA associated with the current submission is available in Github.[1)]

Error Metrics

A qualitative and quantitative analysis has been proposed for evaluating the performance of GANs in arresting data distributions and judging the quality of the generated images.

Qualitative Analysis

To qualitatively evaluate the quality of the images, we compared the generated images to the real images one by one (100 DCGAN-generated images and 1000 generated by StyleGAN2-ADA). Specifically, an orthopedic surgeon analyzed the important features of OA joints in two steps. First, compared disease-specific anatomical characteristics such as the size, number, and location of osteophytes, joint

[1)] https://github.com/gunahn/StyleGAN2-ADA-pytorch-for-gray-images

space narrowing, subchondral bone sclerosis, and subluxation. Second, compared technical image quality factors such as defined edge, soft tissue shadow, homogeneity, blur effects, and color hue.

Visual Turing Test

To evaluate visual fidelity with generalizability, a previous study utilized expert users.[10] To validate the performance of the generated images, we conducted a Visual Turing test. The test involved two orthopedic surgeons, two computer vision experts, and one musculoskeletal radiologist. They were asked to classify 100 randomly mixed images consisting of 50 generated images (generated by StyleGAN2-ADA), and 50 real images, as either real (class 1) or generated (class 0). We informed that distribution will be 50% of real images and 50% of generated ones.

Fréchet Inception Distance (FID)

To quantitatively judge sample realism, Fréchet inception distance (FID) was recently shown to be a sensibly good metric for comparing image distributions. The FID captures the similarity between generated and real images better than the inception score. The FID is known to be consistent with increasing disturbances and human judgment.[23] We calculated all FIDs between 10k real knee images and 100 generated knee images for DCGAN and used average. Also, for StyleGAN2-ADA generated images, 10k real knee images and 10k generated images were used to calculate the average FID.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is feature extraction and data representation technique extensively used in the areas of computer vision.[24] Medical images are usually huge collections of information, thus are difficult to integrate and process,

consuming comprehensive computing time. Therefore, high-dimensional data should be described by a low-dimensional representation.[25] We used PCA to reduce the dimension of generated images by StyleGAN2-ADA and real images to visualize in 2-dimension. Principal components are created in order of the amount of variation they cover: principal component (PC1) captures the most variation, principal component 2 (PC2) captures the second most, and so on. After that, a t-test was conducted in both PC1 and PC2 between generated images by StyleGAN2-ADA and real images. A t-test is a commonly used statistical technique to determine whether the means of two data sets are statistically different from each other.[26]

Finding the Optimal Number of Input Data

In this study, we investigated how many initial input images are required to form reliable gray knee X-rays images with GAN. Our goal is to generate reliable images with as few as possible numbers of the original datasets. Therefore, we changed the number of input knee X-ray images as 1k, 1.5k, 2k, 2.5k, 3k, 5k, 7k, 9k, and 10k by SNUH dataset, trained StyleGAN2-ADA, and all FID was calculated according to the number of augmented data in each network. Moreover, the optimal number of input images was calculated by calculating the slope, which is the change in FID according to the increase in input data.

Ganspace

Härkönen et al. identified important latent directions based on PCA applied either in latent space or feature space[27]. They showed that important directions in GAN latent spaces can be found by applying PCA in latent space for StyleGAN2 and layer-wise decomposition of PCA edit directions leads to many interpretable controls. Gaussian z is sampled N (=10^6) and mapped to W-space (w = M(z)). Then, PCA is performed

with N vectors(w) to obtain a matrix V, a set of principal components. (V is a basis for W) Finally, each control parameter x can be regulated to control the desired principal components to produce a crafted image I'. ($w' = w + Vx$)

Results

Qualitative Analysis

Figure 1. Examples of generated knee X-ray images with the help of a (a) DCGAN (512x512 pixel) and (b) StyleGAN2-ADA (512-512 pixel, input images of 10,000, and augmented images of 800k).

Although the images generated by DCGAN looked realistic, the generated images could easily be identified as generated because of the color hue and outline of the bones (Fig. 1a). In contrast, StyleGAN2-ADA architecture produced high-quality image variation. The StyleGAN2-ADA samples (Fig. 1b) seemed highly realistic, reflecting the characteristics of the outline, osteophytes, joint space, and knee alignment.

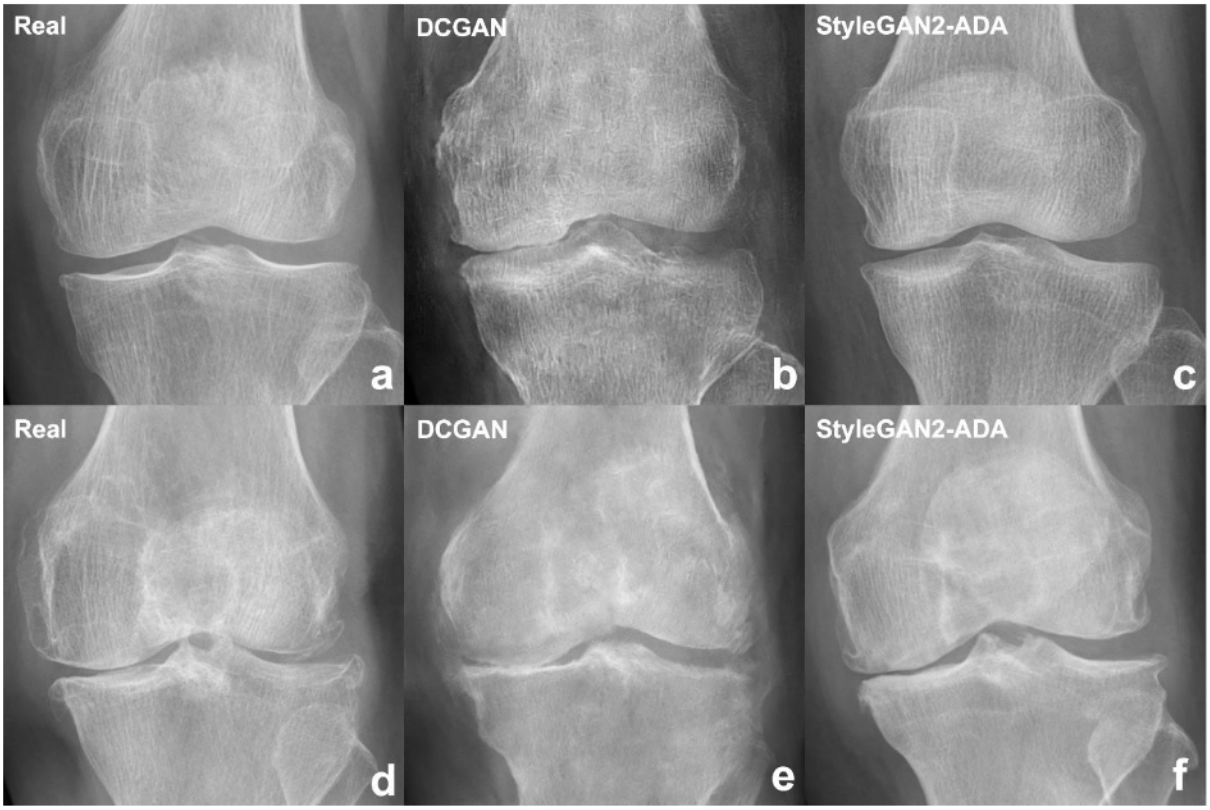

Figure 2. Early osteoarthritis images are shown in a, b, and c which could be compared with technical image quality factors. (b) The images generated by DCGAN showed blurred and irregular bond borders compared to the real knee X-ray images. (c) The synthetic StyleGAN2-ADA images seemed highly realistic. Late osteoarthritis images are shown in d, e, and f which could be compared based on disease-specific anatomical characteristics and technical image quality factors. (e) The images generated by DCGAN showed that the osteophyte was wide and the location of the osteophyte was unusual. Also, the margin of the osteophytes was blurred and overlapped. (f) The synthetic StyleGAN2-ADA images were difficult to distinguish from the real images.

We conducted qualitative analysis between the real knee X-ray and synthetic images generated by DCGAN and StyleGAN2-ADA (Fig. 2). First, based on disease-specific anatomical characteristics, the DCGAN-synthesized image (Fig. 2e) showed

that the osteophyte shape was wider than in the real images and the direction of the osteophyte was unusual. Also, the margin of the osteophytes was blurred and overlapped between the femoral and the tibial side. However, in images synthesized by StyleGAN2-ADA, there was no significant difference compared to the real images. For instance, the image generated by StyleGAN2-ADA (Fig. 2f) is difficult to distinguish because there is an osteophyte in the same position, with the same size, direction, and shape in the real images.

Second, based on technical image quality factors, the images generated by DCGAN showed blurred and irregular bone borders compared to the real knee X-ray images. A discontinuity of the fibular head was seen, and the patella margin was not visible in the images generated by DCGAN (Fig. 2b and 2e). The heterogeneity in the color of the synthetic DCGAN images was well-detected in the soft tissue shadows. We supposed that DCGAN was not capable of discriminating between two structures of different densities adjacent to each other. Unlike the obscurity of DCGAN, the synthetic StyleGAN2-ADA images displayed no significant difference in margins, soft tissue shadows, or blurred effects compared to the real images.

Visual Turing Test

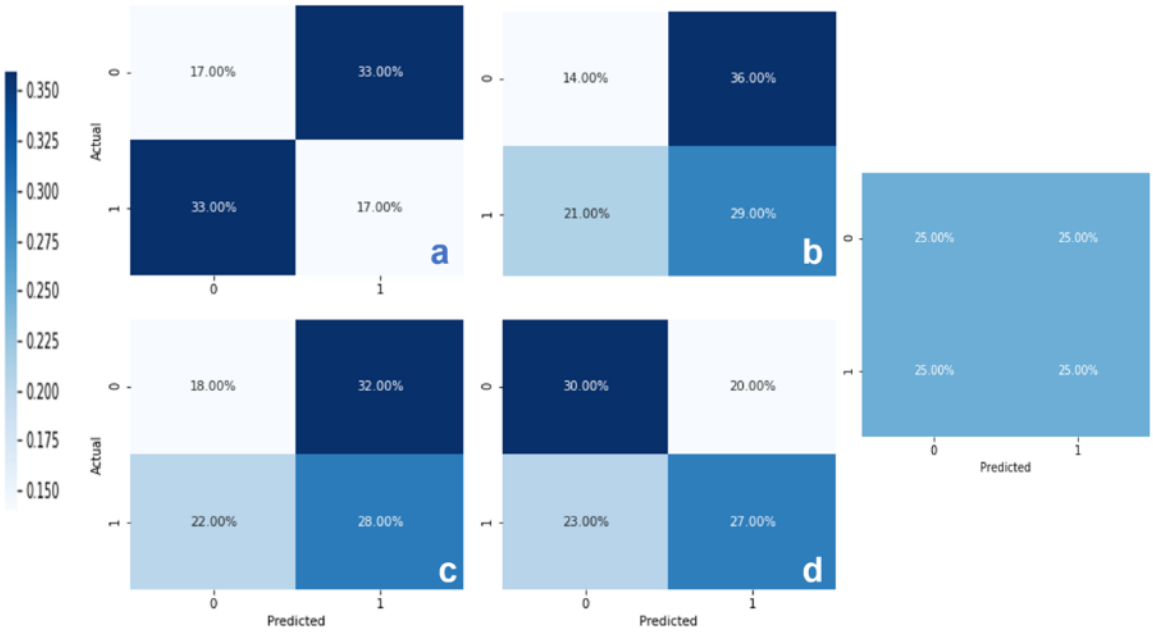

Figure 3. A Visual Turing test was conducted by randomly mixing 100 images consisting of 50 generated images (generated by StyleGAN2-ADA), and 50 real images, either real (class 1) or generated (class 0). Confusion matrix results of the Visual Turing test from orthopedic surgeons (a, b), computer vision experts (c, d), and musculoskeletal radiologists (e) are shown.

The results of the Visual Turing test of all the participants are presented as a confusion matrix in Figure 3. Of note, the classification accuracy from two orthopedic surgeons, two computer vision experts, and one musculoskeletal radiologist are 34, 44, 46, 57, 50 for each, implying that neither the computer vision experts, orthopedic surgeons, nor musculoskeletal radiologists could not reliably distinguish between the real and generated images.

Fréchet Inception Distance

For quantitative comparison, we compared the FID to the real data. FID values compared to DCGAN and real images were shown as 117, as well as FID values

between StyleGAN2-ADA and real images were 2.96. The FID was considerably lower in the StyleGAN2-ADA model, which reflects a visual exploration of the generated images.

PCA & t-test

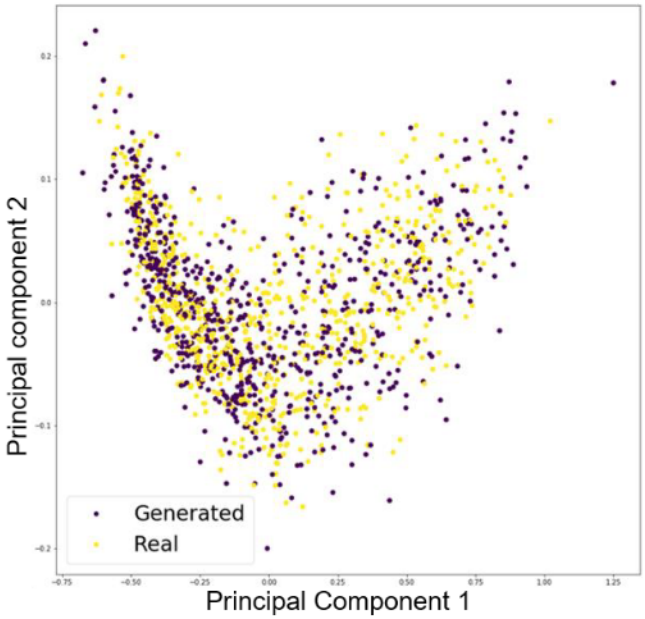

Figure 4. As we can see from the visualization of real and generated images distribution using two-dimensional PCA mapping, there was no visual difference in the distribution of real and generated by StyleGAN2-ADA images.

Table 1. Comparison of principal component analysis between real and generated images by StyleGAN2-ADA

	Mean	STD	t-value	P-value
PC1				
Generated images	-0.01	0.37	-1.26	0.21
Real images	0.01	0.37		
PC2				
Generated images	0	0.06	-0.14	0.88

Real images	0	0.06

PC: principal component, STD: standard deviation

We conducted a 2D visualization of real and generated images by dimension reduction with help of PCA (Fig. 4). There was no visual difference between real and generated image distribution. We used a t-test to evaluate whether AI could distinguish between real and generated images without supervision (Table 1). T-test results showed that a significant difference did not exist between the real and generated images ($P > 0.05$).

Finding the Optimal Number of Input Data

For finding the optimal number of input data, we trained StyleGAN2-ADA with different subsets of SNUH data (Figure 5). Training started a similar way in each case, but eventually, the FID started to diverge. The less training data there is, the faster diverges happen. Complete FID details are presented in Table 2. Slope (ΔFID/Δ Input images) showed more clearly that as the data exceeds 2000, FID does not significantly decrease as the input data increases.

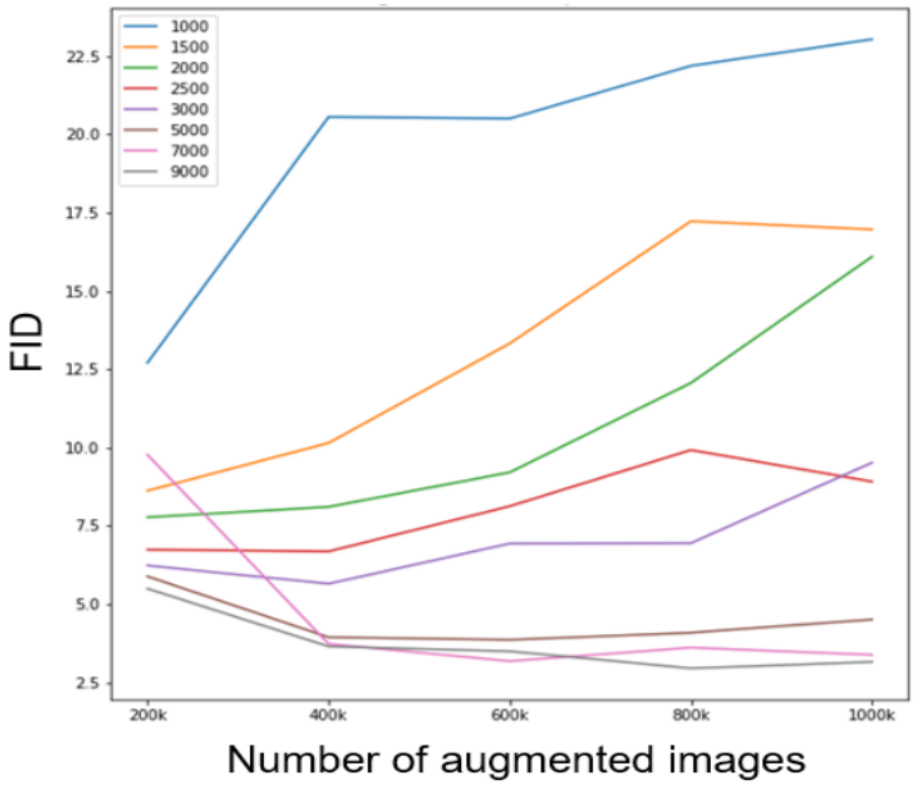

Figure 5. Training curves are shown for knee X-ray images from SNUH with different training set sizes using adaptive augmentation.

Table 2. Error metrics (FID) by input data and augmented data using StyleGAN2-ADA

	Augmented data (K)					ΔFID_max/Δ Input
	200	400	600	800	1000	images (1/K)
Input data (K)						
1	12.7*	20.6	20.5	22.2	23.0	- 12.7
1.5	8.62*	10.1	13.3	17.2	16.9	-8.16
2†	7.77*	8.10	9.21	12.1	16.1	-1.7
2.5	6.74*	6.68	8.13	9.92	8.91	-2.06
3	6.23	5.65*	6.93	6.95	9.52	-2.18
5	5.89	3.94	3.86*	4.08	4.51	-0.895
7	9.77	3.73	3.17*	3.61	3.38	-0.345

9	5.49	3.65	3.49	2.95*	3.15	-0.11
10	5.79	3.78	3.30	2.96*	3.07	0.005

*The value denotes the lowest FID in each network.

†Slope (ΔFID/Δ Input images) showed more clearly that as the data exceeds 2000, FID did not significantly decrease as the input data increased.

FID: Fréchet inception distance

GANspace

Figure 6 showed that generated images can show arthritis progression by controlling latent vectors in PC3. There were no meaningful clinical changes in other principal components. The most striking result to emerge from the data is that unsupervised identification of interpretable directions in an existing GAN was possible. The greater the sigma of latent space in PC3, the worse arthritis becomes. By this, we were able to make diverse datasets including KL-grading 1 to 4 freely. Arthritis progression was shown by joint space narrowing, number, and sizes of osteophyte increases, subchondral bone sclerosis increases, and lateral subluxation increases.

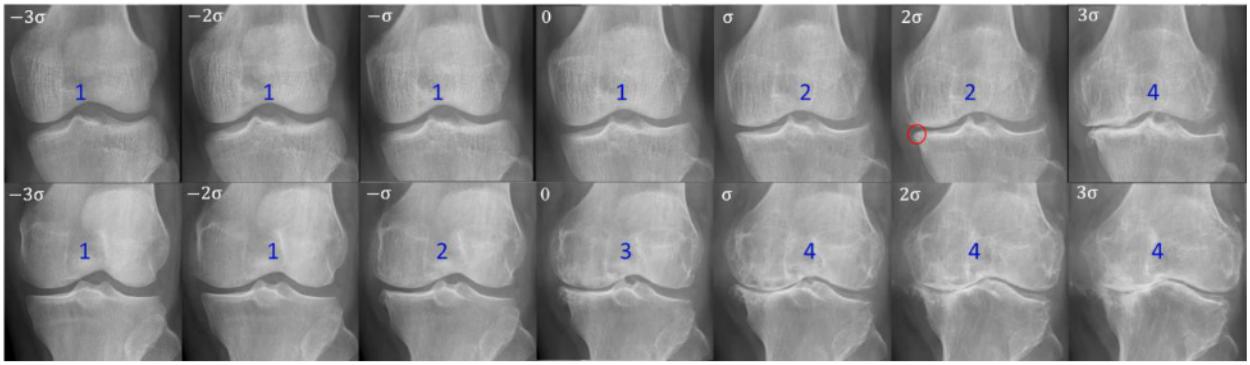

Figure 6. Examples of generated images that can show Arthritis progression by controlling latent vector of PC3. (0 in the middle is the input images and other images are generated by changing sigma value in the latent space, numbers indicate KL-grade)

Discussion

In this study, we obtained a very realistic image by using StyleGAN2-ADA and analyzed generated Knee AP X-ray images qualitatively and quantitatively. This is expected to help use the knee AP synthetical image in the future to make a dataset that contains characters that we want. Overall, our research opens up new possibilities and demonstrates the potential of employing generative models to synthesize realistic images to solve data scarcity and class inequalities.

Other studies have applied GAN on various plain radiographs for various medical applications.[8-11] However, to the best of our knowledge, this is the first study to apply the state-of-the-art GAN method to generate 512×512 images with quality verification. We used StyleGAN2-ADA which allows for the generation of images from a limited number of datasets. Although modern high-quality, high-resolution GAN requires around 1M images, StyleGAN2-ADA generates high-quality, high-resolution images with fewer images by using a wide range of augmentations to prevent the discriminator from overfitting, while ensuring that none of the augmentations leak to the generated images.[20] Overall, the knee X-ray image generation model using our dataset showed as much performance as the other models generated through other databases [6, 28] with only 10K images.

A previous study demonstrated the size and direction of osteophytes in knee plain radiographs and the association with joint space narrowing and subchondral sclerosis.[29] In the DCGAN synthetic images, the osteophyte shape was wider than in the real images and the direction of the osteophyte was unusual. The margins of the osteophytes were blurred and overlapped between the femoral and tibial sides. And some images with clearly visible joint space narrowing were not subchondral sclerosis. In contrast, in the images generated by StyleGAN2-ADA, it was difficult to

distinguish between the real and generated images because the osteophyte was in the same position, size, direction, and shape as the real images.

Of note, the images generated by StyleGAN2-ADA were so realistic that even experts from both the computer vision and medical fields were not able to distinguish between the real images and the generated ones, showing an average classification accuracy of 46%. This is meaningfully lower accuracy than other research that shows a classification accuracy of 58% between real images and generated images by Progressive GAN.[10] Since the computer vision experts were familiar with common GAN artifacts, they were expected to identify generated images, although they did not know about judging actual knee X-ray images. In contrast, orthopedic surgeons know the important features of knee images and possible arthritis progression, even though they may not be familiar with computer vision. On top of that, musculoskeletal radiologists would recognize unnatural features of generated images well since they have been seeing diverse real knee images while interpreting plain radiographs and performing diagnostic and therapeutic joint injection procedures.

From a quantitative analysis perspective, it was difficult to distinguish between real and generated images by computer vision as well. The images generated by StyleGAN2-ADA showed significantly shorter FID (2.96) than those generated by DCGAN (117), consistent with the qualitative analysis. If we compare our FID results with those of other datasets, images with a resolution of 512×512-pixel generated for BreCaHAD[30] (1944 images) showed 15.71, AFHQ-CAT (5153 images) showed 3.55, AFHQ-Dog(4739 images) showed 7.40, and AFHQ-WILD[31] (4738 images) showed 3.05 FID. Furthermore, the t-test result illustrated that a statistical difference did not exist in PCA between the real and generated images by StlyeGAN2-ADA ($P > 0.05$).

Overall, we can conclude that the synthetic samples were indeed highly realistic.

One of the purposes of this study was to identify conditions such as the number of training images required to attain reliable generative models. Although GAN is used to obtain reliable data relatively easily, the performance of GAN itself is also closely related to how much input data it contains.[22] The key problem with small datasets is that the discriminator overfits the training examples. [32] In small datasets feedback to the generator becomes meaningless and training starts to diverge. Medical data is particularly costly and time-consuming to integrate and standardize in one place compared to other data. Since adding input data after more than 2,000 data did not significantly improve performance in our result, we concluded that about 2,000 was the most appropriate optimal input size for generating knee radiographs images.

Furthermore, we showed the feasibility of using the GAN method that can overcome the data imbalance problem. The single most striking observation to emerge from the result was that AI found the axis that can represent a progression of arthritis in an unsupervised manner. It is worth noting that AI found the axis of arthritis among the various axes that show the change of knee by just putting in images of various knees without giving any information on the progress stage of arthritis. This is expected to help use the knee AP image synthetically in the future to make a dataset that contains characters that we want, such as diverse KL grades. This dataset is expected to play a major role in the future as it can predict the progression of arthritis and identify patients who need to be treated more aggressively. The present study raises the possibility that the GAN model can be used to show a future progression of arthritis in a clinical sense.

To sum up, our research opens up new possibilities and demonstrates the

potential of employing generative models to synthesize realistic images to solve data scarcity and class inequalities. This study will allow tens of thousands of image data to be generated quickly and inexpensively and replace actual medical data containing sensitive information and facilitate saving time and money for data labeling operations and secure accurate labeled datasets.

Limitations and Future work

First, in some real knee X-rays, the fabella was detected behind the lateral femoral condyle. The fabella usually overlaps in X-ray images. And not everybody has fabella. Therefore, there were only a few images that we could detect fabella (Fig. 6a). However, in the synthetic images, there was no evidence of the fabella. Some images generated by StyleGAN2-ADA showed a fabella-like bony shadow around the lateral femoral epicondyle (Fig. 6b). We concluded that this was one of the unrealistic characteristics of the synthetic images.

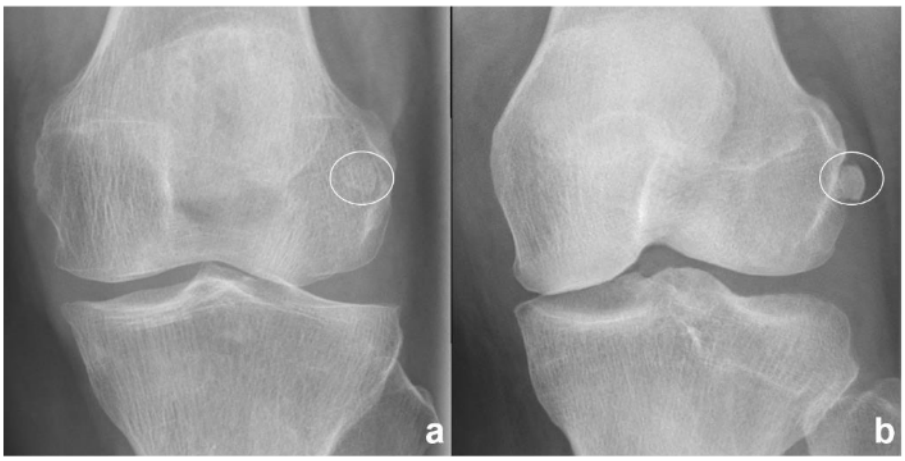

Figure 7. A real knee X-ray is shown in a and an image generated by StyleGAN2-ADA is shown in b. The circle in image a is the fabella behind the lateral femoral condyle. Image b shows a fabella-like bone shadow around the lateral femoral epicondyle.

Since the performance of the generated images significantly relies upon the number of input images,[33] we could use the generated images as new input for GANs. Also, in a previous study, a researcher improved the quality of generated images by extracting face semantic maps as well as head pose angles, glasses type, and eye occlusion score.[34] Similarly, we can try to improve the quality of images by extracting knee semantic maps such as osteophyte, joint space, and sclerosis. By doing that, we expect that a broken tibia or the odd edge of the knee would not be observed.

Furthermore, by algorithmic improvement, GANs could be trained effectively using even smaller datasets. If there are more than 10K images, GAN is not necessary. However, if a GAN could be trained reliably on small datasets like 1K or 0.5K, the GAN could revolutionize medical research for rare diseases by overcoming imbalance problems. Also, medical image labeling is a very demanding task, so if we can freely generate data by class using GANs, it will reduce the burden on many researchers in the future.

Acknowledgment

We would like to extend our sincere thanks to the experts who were willing to conduct the Visual Turing test.

Conflict of Interest

All authors declare that there are no conflicts of interest.

References

1. Marcu, L. G., Boyd, C. & Bezak, E. Current issues regarding artificial intelligence in cancer and health care. Implications for medical physicists and

biomedical engineers. *Health and Technology* vol. 9 375–381 (2019).

2. Zhang, J., Xie, Y., Wu, Q. & Xia, Y. Medical image classification using synergic deep learning. *Med. Image Anal.* **54**, 10–19 (2019).
3. Kayaalp, M. Patient Privacy in the Era of Big Data. *Balkan Med. J.* **35**, 8–17 (2018).
4. Mooney, S. J. & Pejaver, V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu. Rev. Public Health* **39**, 95–112 (2018).
5. Kramberger, T. & Potočnik, B. LSUN-Stanford Car Dataset: Enhancing Large-Scale Car Image Datasets Using Deep Learning for Usage in GAN Training. *Applied Sciences* vol. 10 4913 (2020).
6. Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**, (2020).
7. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **58**, 101552 (2019).
8. Wolterink, J. M., Leiner, T., Viergever, M. A. & Isgum, I. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. *IEEE Trans. Med. Imaging* **36**, 2536–2545 (2017).
9. Frid-Adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* vol. 321 321–331 (2018).
10. Baur, C., Albarqouni, S. & Navab, N. Generating Highly Realistic Images of Skin Lesions with GANs. *Lecture Notes in Computer Science* 260–267 (2018) doi:10.1007/978-3-030-01201-4_28.
11. Bermudez, C. *et al.* Learning Implicit Brain MRI Manifolds with Deep Learning.

Proc. SPIE Int. Soc. Opt. Eng. **10574**, (2018).

12. Lubis, A. M. T., Wonggokusuma, E. & Marsetio, A. F. Intra-articular Recombinant Human Growth Hormone Injection Compared with Hyaluronic Acid and Placebo for an Osteoarthritis Model of New Zealand Rabbits. *Knee Surg. Relat. Res.* **31**, 44–53 (2019).
13. Navarro, R. A. *et al.* Does Knee Arthroscopy for Treatment of Meniscal Damage with Osteoarthritis Delay Knee Replacement Compared to Physical Therapy Alone? *Clin. Orthop. Surg.* **12**, 304–311 (2020).
14. Sim, J. A. *et al.* Utility of preoperative distractive stress radiograph for beginners to extent of medial release in total knee arthroplasty. *Clin. Orthop. Surg.* **1**, 110–113 (2009).
15. Moon, Y. W. *et al.* Factors correlated with the reducibility of varus deformity in knee osteoarthritis: an analysis using navigation guided TKA. *Clin. Orthop. Surg.* **5**, 36–43 (2013).
16. Altman, R. D. & Gold, G. E. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and Cartilage* vol. 15 A1–A56 (2007).
17. Kellgren, J. H. & Lawrence, J. S. Rheumatoid Arthritis in a Population Sample. *Annals of the Rheumatic Diseases* vol. 15 1–11 (1956).
18. Choi, E.-S. *et al.* Relationship of Bone Mineral Density and Knee Osteoarthritis (Kellgren-Lawrence Grade): Fifth Korea National Health and Nutrition Examination Survey. *Clin. Orthop. Surg.* **13**, 60–66 (2021).
19. Bowes, M. A. *et al.* Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative. *Ann. Rheum. Dis.* (2020) doi:10.1136/annrheumdis-2020-217160.
20. Carlson, V. R. *et al.* Compliance With the AAOS Guidelines for Treatment of

Osteoarthritis of the Knee: A Survey of the American Association of Hip and Knee Surgeons. *J. Am. Acad. Orthop. Surg.* **26**, 103–107 (2018).

21. Goodfellow, I. *et al.* Generative adversarial networks. *Communications of the ACM* vol. 63 139–144 (2020).
22. Karras, Tero, et al. "Training generative adversarial networks with limited data." Advances in Neural Information Processing Systems 33: 12104-12114 (2020).
23. Obukhov, A. & Krasnyanskiy, M. Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance. *Software Engineering Perspectives in Intelligent Systems* 102–114 (2020) doi:10.1007/978-3-030-63322-6_8.
24. Peng, Y. *et al. Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part III*. (Springer Nature, 2020).
25. Kaya, I. E., Pehlivanlı, A. Ç., Sekizkardeş, E. G. & Ibrikci, T. PCA based clustering for brain tumor segmentation of T1w MRI images. *Comput. Methods Programs Biomed.* **140**, 19–28 (2017).
26. Kim, T. K. T test as a parametric statistic. *Korean Journal of Anesthesiology* vol. 68 540 (2015).
27. Härkönen, E., Hertzmann, A., Lehtinen, J. & Paris, S. GANSpace: Discovering Interpretable GAN Controls. *Adv. Neural Inf. Process. Syst.* **33**, 9841–9850 (2020).
28. Karras, T. *et al.* Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020) doi:10.1109/cvpr42600.2020.00813.
29. Nagaosa, Y. Characterisation of size and direction of osteophyte in knee

osteoarthritis: a radiographic study. *Annals of the Rheumatic Diseases* vol. 61 319–324 (2002).

30. Aksac, A., Demetrick, D. J., Ozyer, T. & Alhajj, R. BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Res. Notes* **12**, 82 (2019).
31. Choi, Y., Uh, Y., Yoo, J. & Ha, J.-W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020) doi:10.1109/cvpr42600.2020.00821.
32. Yazici, Y., Foo, C.-S., Winkler, S., Yap, K.-H. & Chandrasekhar, V. Empirical Analysis Of Overfitting And Mode Drop In Gan Training. *2020 IEEE International Conference on Image Processing (ICIP)* (2020) doi:10.1109/icip40778.2020.9191083.
33. Shmelkov, K., Schmid, C. & Alahari, K. How Good Is My GAN? *Computer Vision – ECCV 2018* 218–234 (2018) doi:10.1007/978-3-030-01216-8_14.
34. Or-El, R., Sengupta, S., Fried, O., Shechtman, E. & Kemelmacher-Shlizerman, I. Lifespan Age Transformation Synthesis. *Computer Vision – ECCV 2020* 739–755 (2020) doi:10.1007/978-3-030-58539-6_44.