Received April 12, 2022, accepted May 5, 2022, date of publication May 16, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175506

Hierarchical Knee Image Synthesis Framework for Generative Adversarial Network: Data From the Osteoarthritis Initiative

HONG-SENG GAN[1], **(Senior Member, IEEE), MUHAMMAD HANIF RAMLEE**[2], **BANDER ALI SALEH AL-RIMY**[3], **(Senior Member, IEEE), YENG-SENG LEE**[4], **AND PRAYOOT AKKARAEKTHALIN**[5]

[1]Department of Data Science, Universiti Malaysia Kelantan, UMK City Campus, Pengkalan Chepa, Kelantan 16100, Malaysia
[2]Bioinspired Devices and Tissue Engineering (BIOINSPIRA) Group, Faculty of Engineering, School of Biomedical Engineering and Health Sciences, Universiti Teknologi Malaysia, Johor Bahru, Johor 81310, Malaysia
[3]Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor 81310, Malaysia
[4]Department of Electronic Engineering Technology, Faculty of Engineering Technology, Universiti Malaysia Perlis, Arau, Perlis 02600, Malaysia
[5]Department of Electrical and Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

Corresponding authors: Hong-Seng Gan (hongseng1008@gmail.com) and Prayoot Akkaraekthalin (prayoot.a@eng.kmutnb.ac.th)

This work was supported in part by Geran Penyelidikan UMK Fundamental (UMK Fund) (Project Title: Attention-based Generative Adversarial Network for Realistic Knee Image Synthesis) through Universiti Malaysia Kelantan under Grant R/FUND/A1500/01934A/2022/01043; in part by the National Science, Research and Innovation Fund (NSRF); and in part by the King Mongkut's University of Technology North Bangkok under Contract KMUTNB-FF-66-10.

ABSTRACT Medical images synthesis is useful to address persistent issues such as the lack of training data diversity and inflexibility of traditional data augmentation faced by medical image analysis researchers when developing their deep learning models. Generative adversarial network (GAN) can generate realistic image to overcome the abovementioned problems. We proposed a GAN model with hierarchical framework (HieGAN) to generate high-quality synthetic knee images as a prerequisite to enable effective training data augmentation for deep learning applications. During the training, the proposed framework embraced attention mechanism before the 256×256 scale in generator and discriminator to capture salient information of knee images. Then, a novel pixelwise-spectral normalization configuration was implemented to stabilize the training performance of HieGAN. We evaluated the proposed HieGAN on large scale knee image dataset by using Am Score and Mode Score. The results showed that HieGAN outperformed all relevant state-of-art. Hence, HieGAN can potentially serve as an important milestone to promote future development of more robust deep learning models for knee image segmentation. Future works should extend the image synthesis evaluation to clinical-related Visual Turing Test and synthetic data augmentation for deep learning segmentation task.

INDEX TERMS Generative adversarial network, knee, image synthesis, biomedical image processing.

I. INTRODUCTION

Existing supervised learning methods in medical image segmentation depend heavily on large quantity of high-quality training data. The problem becomes apparent with the resurgence of deep learning, whose training requires huge volume of labelled data. To build-up large scale training datasets represent a daunting task for most medical image analysis researchers given the enormous financial costs and expert label time involved. Meanwhile, traditional augmentation methods such as scaling, rotating, flipping and elastic deformation fail to consider the variations in size, shape, location and appearance of specific pathology [1].

Generative Adversarial Network (GAN) [2] is a powerful unsupervised training approach. It learns the pattern of input samples and generate new outputs based on underlying structural information in training data. As a result, GANs are very useful for medical image synthesis. Prospectively, synthetic medical image-based augmentation offers solution to the lack of manually annotated data and inflexibility in traditional

The associate editor coordinating the review of this manuscript and approving it for publication was R. K. Tripathy.

augmentation. Moreover, synthetic medical images are not associated with individual patient image information. There is no concern on data privacy regulations when sharing the data for reproducibility purpose.

Generation of high-quality synthetic medical image is a prerequisite to numerous image processing applications, including segmentation. In recent years, numerous works related to synthetic images generation using GANs have been found. Unconditional synthesis produces an image from the latent space of real image without any conditional information [3] and are often trained with more than 10,000 images [4]. For example, Sketching-rendering Unconditional GAN (SkrGAN) has been proposed to generate several types of synthetic medical images, including retinal color fundus, chest x-ray, lung computed tomography (CT) and brain magnetic resonance image (MRI) [5]. Nevertheless, GANs are notorious for their instable training performance. Deep Convolutional GAN (DCGAN) [6] and Progressive Growing GAN (PGGAN) [7] are two widely adopted in medical image synthesis attributed to better training stability.

Chuquicusma *et al.* (2018) used DCGAN to generate benign and malignant lung nodule samples of size 56×56. Their Visual Turing Test results concluded that DCGAN produced highly realistic class specific nodules but the inter-observer error was relatively high [8]. Frid-Adar *et al.* (2018) applied three DCGANs to generate CT images for three classes of liver lesion i.e., cysts, metastases, and hemangiomas. The image size was 64×64. Accordingly, synthetic lesion images were found to be beneficial to classification task with improved sensitivity and specificity when combined with real training data [9]. Both studies concentrated on augmenting the training data size of different pathology classes and the capability of DCGAN is limited to low-resolution medical images.

PGGAN is capable to generate synthetic medical image of higher resolution up to 1024×1024 [7]. Beers *et al.* (2018) generated synthetic retinal fundus image and MR image of brain up to 512×512 to protrude subtle pathological features at the expense of heavy computational cost and slow training speed [10]. To date, PGGAN has been generalized to various image types such as gastritis image [11], cardiac MR image [12], body CT images [13], mammograms [14] and chest X-ray image [15]. In most works, the anatomical variation of organ is either small or cropped to the region of interest (ROI) before training. Otherwise, pertinent features from medical image fail to be synthesized.

Besides, it is reported that the performance of PGGAN on mode collapse problem remains under expectation. Another work focuses on improving the training stability of GAN is illustrated through the introduction of Wasserstein GAN (WGAN) [16]. Wasserstein distance is utilized to replace conventional Jensen-Shannon Divergence in the training objective. While the model training becomes more stable, it is challenging to enforce the Lipschitz constant and hard to recognize complex image landscape. Some GANs are developed to tackle the salient feature recognition during the image synthesis process. For example, Auto-Embedding GAN (AEGAN) [17] has been proposed to generate high resolution images. The model learns a latent embedding extracted from an autoencoder. The model has reported good results for image synthesis task by using a single TITAN X Pascal GPU.

Human knee structure comprises of multiple types of musculoskeletal tissues [18]. As illustrated in Fig.1, there are three different cartilages and corresponding bone components that form the overall knee structure. Their anatomical geometries change significantly across the image slices [19]. An effective medical image synthesis framework that can maintain training stability and capture the fine details of irregular knee structure posts a great challenge that has never been addressed before. Besides, existing GANs [7], [20], [21] are mainly applied on CIFAR-10 and/or CIFAR-100 datasets. Natural image-based evaluation metrices such as Inception Score and Fréchet Inception Distance are used in the works. Hence, a novel knee image synthesis via hierarchical framework is proposed. Specifically, main contributions of this paper include:

1. The proposed knee image synthesis model in hierarchical architecture composes of layer-by-layer training structure with attention added in between the layers is designed to enhance its salient feature recognition capability.
2. Training stability of the proposed framework is improved by adopting a hybrid pixelwise-spectral normalization configuration in generator and discriminator in order to avoid incurring additional computational cost to the model training process.
3. Instead of Inception Score and Fréchet Inception Distance applied for most natural image evaluation, distance-based Wasserstein Distance and Mean Absolute Error as well as probability-based Mode Score and AM Score are adopted to better evaluate the proposed HieGAN framework in the context of medical image synthesis.
4. Findings have suggested that the proposed HieGAN framework can serve as promising medical image data augmentation option to tackle the scarcity of training data and class imbalance problem faced by existing deep learning segmentation models.

II. MATERIALS AND METHODS

A. IMAGE DATASETS

The study comprised of 75 normal knee image datasets. MR image data was acquired by using 3.0 Tesla (T) MRI Scanner (Siemens Magnetom Trio, Erlangen, Germany) with quadrature transmit-receive knee coil (USA Instruments, Aurora, OH). Dual echo steady state (DESS) with water excitation (WE) imaging sequence was selected [22]. All knee image datasets were chosen randomly from the Osteoarthritis Initiative (OAI) database. The images have section thickness of 0.7 mm and an in-plane resolution of 0.365×0.365 mm^2

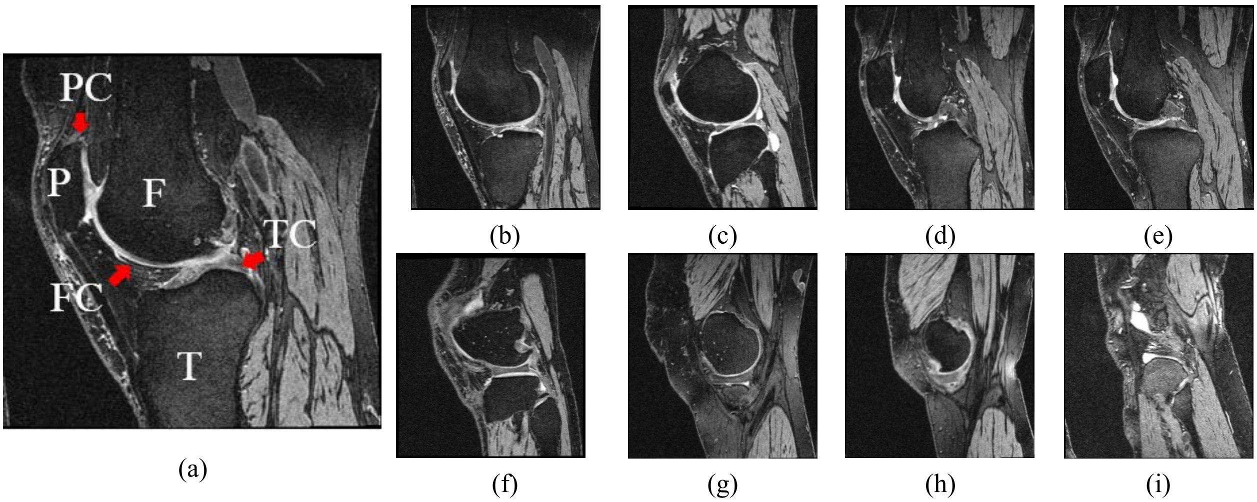

FIGURE 1. Knee structure with Patella (P), Femur (F), Tibia (T), Patellar cartilage (PC), Femoral cartilage (FC) and Tibial cartilage (TC) (a). Knee structure components such as knee bone, cartilage, muscles and ligaments have changing anatomical geometry as illustrated in (b)–(i).

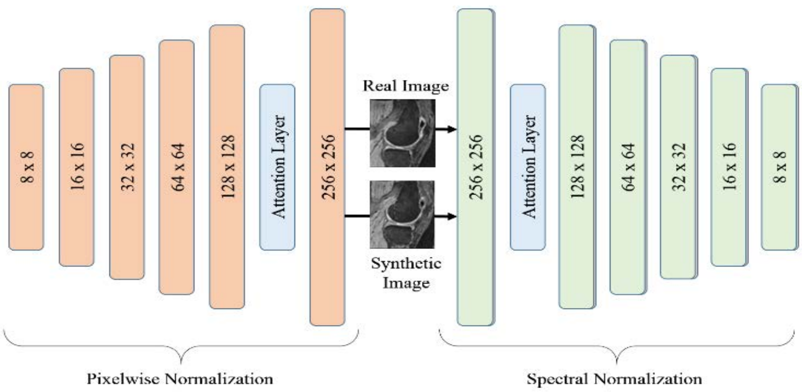

FIGURE 2. Overview of the HieGAN with hierarchical architecture, which comprises of attention and hybrid normalization configuration.

(field of view = 140 × 140 mm, flip angle = 25°, TR/TE = 16.3/4.7 msec, matrix size = 384 × 384mm, bandwidth = 185 Hz/pixel). More details about the dataset can be found at http://oai.epi-ucsf.org/datarelease/About.asp.

B. ARCHITECTURE OF HieGAN

GAN architecture comprises of a generator (G) and a discriminator (D). The generator is responsible to produce synthetic image with distribution indistinguishable from training distribution while the discriminator is trained to distinguish between true samples and fake samples produced by the generator. Both D and G continuously engage in a min-max game given as

$$\min_{G}\max_{D} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}\left[logD(x)\right] + \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}\left[log\left(1-D\left(\tilde{x}\right)\right)\right] \quad (1)$$

where $\mathbb{P}_r$ is the data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $\tilde{x} = G(z), z \sim p(z)$. The input z to the generator is sampled from noise distribution p. Simultaneous training between these two competing components contribute to uneasy convergence or failure modes when equilibrium cannot be achieved.

Intuitively, the HieGAN model trains progressively from low-resolution (8 × 8) to high-resolution (256 × 256) scale. The model learns the diverse spatial features of knee images. As the HieGAN progresses, local spatial features are acquired in higher resolution layers. During the transition, nearest neighbor upsampling method is used to fade the new scale in. This approach helps to overcome sudden shocks to the existing trained, lower resolution scales. We also discover that Minibatch discriminator is computationally complex and sensitive to hyperparameter selection; so, we have selected a pixelwise-spectral normalization configuration.

Knee structure exhibits constant change of shape and size within the dataset. Conventional convolution operation might fail to capture the diversity of variation during model training, which can compromise the quality of image. An attention layer is proposed and implemented before the 256 × 256 scale in generator and discriminator. The architecture of HieGAN is illustrated in Fig. 2.

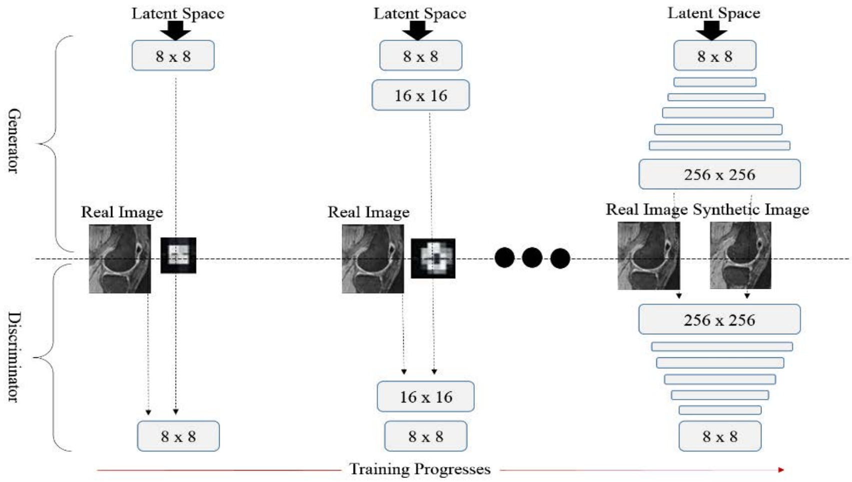

FIGURE 3. Training flow of HieGAN framework started from 8 × 8 and double-up until 256 × 256. At each scale, the model is trained until convergence is attained.

C. TRAINING OF HieGAN

In Fig. 3, training of HieGAN started from 8 × 8 and trained progressively until 256 × 256. A total of 14,920 training images were used to train the HieGAN. The model was implemented by using Tensorflow in Python. During the training, we set the learning rate, α at 0.001, maximum iteration at 228,000, input noise at 256 and noise standard deviation at 0.01 in generator and discriminator. Leaky RELU was used in the model. The batch size at 8 × 8 is 128, 16 × 16 is 64, 32 × 32 is 32, 64 × 64 is 16, 128 × 128 is 8 and 256 × 256 is 4. For optimization, Adam was used with $\beta_1 = 0$, $\beta_2 = 0.999$ and $\varepsilon = 1 \times 10^{-8}$. All training was performed on desktop with NVIDIA GeForce RTX 3070 GPU. The model training took 3 weeks.

Original WGAN [16] utilizes weight clipping to attain 1-Lipschitz functions. But the weight clipping is susceptible to optimization difficulties, capacity underuse and exploding/vanishing gradient without careful tuning of the weight clipping parameter c. We adopted Wasserstein loss function plus gradient norm penalty to achieve Lipschitz continuity in discriminator. The loss function was first proposed in WGAN with Gradient Penalty (WGAN-GP) [23]. A gradient penalty is a soft version of the Lipschitz constraint, which follows from the fact that functions are 1-Lipschitz if and only if the gradients are of norm at most 1 everywhere. The squared difference from norm 1 is used as the gradient penalty as shown in

$$L_D = \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}\left[D(\tilde{x})\right] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}\left[D(x)\right] + \underbrace{\lambda \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}\left[(\|\nabla_{\tilde{x}}D(\tilde{x})\|_2 - 1)^2\right]}_{Gradient\ Penalty} \quad (2)$$

The generator loss function remains unchanged as shown in

$$L_G = D(\tilde{x}) \quad (3)$$

Pixelwise normalization was initially implemented in [7] to normalize every feature vector in pixel to unit length after each convolutional layer and avoid magnitudes in generator from spiraling out of control as a result of competition with discriminator. The formulation is illustrated in

$$b_{x,y} = \frac{a_{x,y}}{\sqrt{\frac{1}{N}\sum_{j=0}^{N-1}\left(a_{x,y}^{j}\right)^2 + \epsilon}} \quad (4)$$

where $\epsilon = 10^{-8}$, N is the number of feature maps and $a_{x,y}$ and $b_{x,y}$ are the original and normalized feature vector in pixel (x, y), respectively.

In this work, the pixelwise normalization is initiated by taking the pixel value of each channel of image at position (x, y). A feature vector for each (x, y), $a_{x,y}$ is constructed and calculates the value for each feature map. Then, each vector is normalized between 0 and 1 by using Eqn. 4. The feature vectors are forwarded to the next layer.

Spectral normalization was proposed in [24] to stabilize the training of discriminator through tuning the hyperparameters. It controls the Lipschitz constant of discriminator f by constraining the spectral norm of each layer $g : h_{in} \rightarrow h_{out}$. The Lipschitz norm $\|g\|_{Lip}$ is equal to $sup_h\sigma(\nabla g(h))$, where $\sigma(a)$ is the spectral norm of matrix A(L_2 matrix norm of A)

$$\sigma(a) = \max_{h:h\neq 0}\frac{\|Ah\|_2}{\|h\|_2} = \max_{\|h\|_2\leq 1}\|Ah\|_2 \quad (5)$$

which is equivalent to the largest singular value of A. Therefore, for a linear layer $g(h) = Wh$ the norm is given by $\|g\|_{Lip} = sup_h\sigma(\nabla g(h)) = sup_h\sigma(W) = \sigma(W)$. Spectral normalization normalizes the spectral norm of weight matrix W so it satisfies the Lipschitz constraint $\sigma(W) = 1$

$$\bar{W}_{SN}(W) = \frac{W}{\sigma(W)} \quad (6)$$

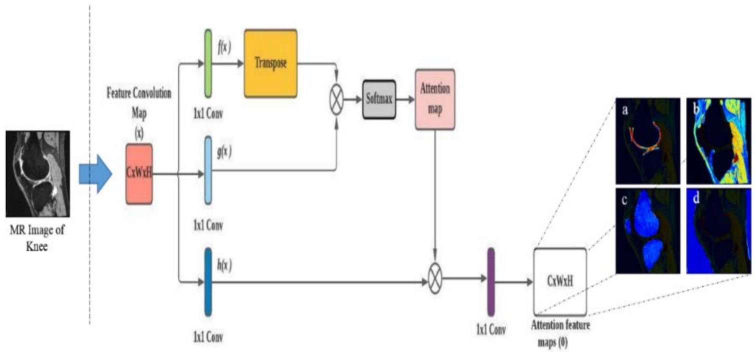

FIGURE 4. Diagram of attention mechanism within the HieGAN architecture. The attention feature convolution map is a result of matrix multiplication between the attention map, $\beta_{j,i}$ and third feature space, $h(x)$ and 1×1 convolution filter. Attention feature m.

D. FINE FEATURE LEARNING VIA ATTENTION

Convolution operator in GANs uses local receptive field to learn local neighborhood representations. The operation curtails effective model learning when specific details are located at different locations. The long-range dependencies can only be processed after passing through several convolutional layers. As a result, PGGAN lacks the power in specifying the features of synthetic knee image. An attention layer computes response at a position as a weighted sum of the features at all positions, where the weights are calculated with only a small computational cost. The mechanism leverages on complementary features in distant portions of the image rather than local regions of fixed shape to generate consistent objects. Thus, it will filter the feature response to retain only the relevant activation.

An attention layer based on non-local model [25] was implemented into the HieGAN architecture. The attention mechanism is exhibited in Fig. 4. Three feature spaces i.e., f, g and h are obtained by using 1×1 convolution. The generator can extract fine details at every location that are carefully coordinated with fine details in distant portions of the knee image while the discriminator can enforce complex geometric constraints on the global image structure. Knee image features from previous hidden layer, $x \in \mathbb{R}^{C \times N}$ are first convolved with 1×1 convolution filter into two feature spaces f and g, where $f(x) = W_f x$, $g(x) = W_g x$, C is the number of channels and N is the number of feature locations of features from the previous hidden layer. $W_f \in \mathbb{R}^{\bar{C} \times C}$ and $W_g \in \mathbb{R}^{\bar{C} \times C}$ are the learned weight matrices.

The feature spaces f and g are linearly combined using a matrix multiplication into $s_{i,j} = f(x_i)^T g(x_j)$. Then, it is fed into the softmax layer to compute the attention to which the model attends to the i^{th} location when synthesizing the j^{th} region. The resultant attention map is given in

$$\beta_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N} exp\left(s_{i,j}\right)} \tag{7}$$

where N is the number of feature maps. Feature vectors of f and g have different dimensions than feature vector h. We multiply the resultant attention map and the third feature space $h(x) = W_h x$ and then convolve with 1×1 convolution filter, $v(x) = W_v x$ to obtain the output of attention layer given in

$$o_j = v\left(\sum_{i=1}^{N} \beta_{j,i} h(x_i)\right) \tag{8}$$

For instance, W_f, W_g, and W_h are the weight matrices of the 1×1 convolutional layer. To enable the generator to learn the local dependence and long-range global dependency of the knee image, we have multiplied the output of attention layer o_j with a weight coefficient, γ and add it to the input feature map x_i to obtain the final output of the attention mechanism given as

$$y_i = \gamma o_j + x_i \tag{9}$$

III. RESULTS

A. EXPERIMENTAL SETTINGS

In this experiment, the HieGAN framework was compared with relevant state-of-art WGAN-GP [23], PGGAN and AEGAN. A total of 5,100 test images were used. Inception score (IS) [26] was the benchmark evaluation metric to gauge the realism of synthetic image sample x produced by GAN based on the Kullback-Leibler (KL) divergence between conditional label distribution, $p(y\,|\,x)$ of samples and marginal distribution, $p(y) \approx \int p(y\,|\,x = G(z))\,dz$ of all samples. The formula is expressed in

$$\begin{aligned} IS &= exp\left(\mathbb{E}_x KL\left(p\left(y\,|\,x\right) \| \, p(y)\right)\right) \\ &= exp\left(H\left(y\right) - \mathbb{E}_x\left[H\left(y\,|\,x\right)\right]\right) \end{aligned} \tag{10}$$

Intuitively, samples are expected to have low entropy, H, if all classes are equally represented in the set of samples (high diversity) and low entropy for easily classifiable samples (better sample quality).

Despite IS has reported well correlation to human evaluation, it is pretrained on GoogleNet. The metric is limited to evaluate training data available at ImageNet classes

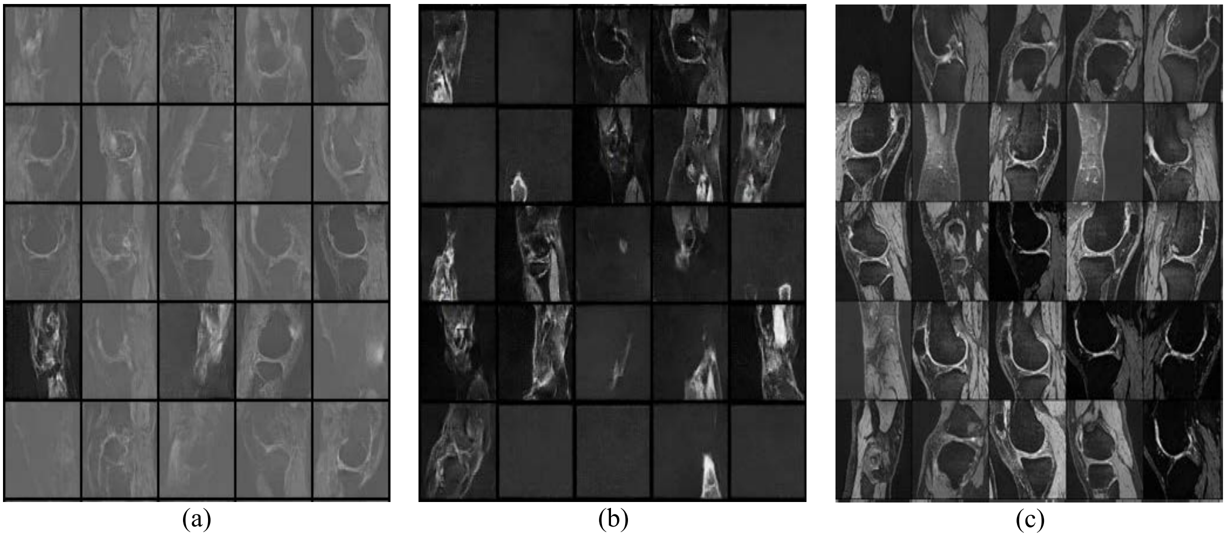

FIGURE 5. Synthetic knee images generated by HieGAN by using different normalization technique configuration in generator (G) and discriminator (D). (a) G: Spectral; D: Spectral, (b) G: Pixelwise; D: Pixelwise, (c) G: Spectral; D: Pixelwise.

instead of medical images. Furthermore, IS fails to recognize images that are generated in mode collapse problem. Similar problems are also reported in Fréchet Inception Distance (FID) [27]. Mode Score (MS) [28] overcomes the limitation of IS by including the prior distribution of label over real data while AM score (AM) [29] incorporates the training data into account by replacing $H(y)$ with KL divergence between y^* and y to address the uneven data distribution problem.

The formulas of MS and AM are expressed in

$$MS = exp\left(\mathbb{E}_x KL\left(p\left(y\,|\,x\right)\|\,p(y^*)\right) - KL(p\left(y\right)\|\,p\left(y^*\right))\right) \tag{11}$$

where x denotes the image data sample and $p(y^*)$ denotes the empirical distribution of labels from training data, y^*. MS has a range between 0 and infinity. Higher MS score reflects better image quality.

$$AM = KL\left(p\left(y^*\right)\|\,p\left(y\right)\right) + \mathbb{E}_x\left(H\left(y\,|\,x\right)\right) \tag{12}$$

AM is minimized when y^* is close to y and entropy of the predicted class label for sample x, $H(y\,|\,x)$ is low. Smaller AM score indicates better image quality. We applied both MS and AM to assess the quality and variation of synthetic knee image. VGG-16 was used as the classifier.

Mean Absolute Error (MAE) and Wasserstein Distance (WD) use distance formulas to compare the probability distributions between real and synthetic data. MAE measures the average magnitude of error between original data distribution, y and synthetic data distribution, $\hat{y}$ over N image samples. The formula is expressed in

$$MAE = \frac{1}{N}\sum_{m=1}^{N}\left|y_m - \hat{y}_m\right| \tag{13}$$

When the synthetic knee image is highly similar to original knee image, the value of MAE will be small.

WD is a measure of distance between probability distribution of synthetic, F and real knee images, G in the feature space of a trained classifier, Inception-v3 network. The formula is expressed in

$$WD = \left(\int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right|^p du\right)^{1/p} \tag{14}$$

where we set the default value parameter $p = 1$ in this work.

B. EVALUATION OF MODEL TRAINING PERFORMANCE

Due to its adversarial nature, training behavior of GAN is always volatile. We have tested different configurations of normalization techniques applied in generators and discriminators of HieGAN. In Fig. 5, we depict the effect of different configuration of normalization techniques on synthesized knee images. In order to illustrate the stability in progressive training due to the use of hybrid normalization configuration, the training performance of HieGAN (refer to Fig. 7) was compared to the PGGAN (refer to Fig. 6) at different scales.

In each plot, training losses of generator and discriminator are computed. Although progressive training has brought stability to training performance, mode collapse was detected at many occasions in PGGAN. The HieGAN also partially suffered from mode collapse but the occurrence was lesser and the training was more stable. Eventually, HieGAN was able to converge in 256 × 256 scale, which has proven it to be more reliable than state-of-art PGGAN.

C. EVALUATION OF SYNTHETIC IMAGE QUALITY

Assessment on the realism between real and synthetic knee images were shown in Table 1. The results suggested that the GANs were improving via training from 8 × 8 until 256 × 256. HieGAN has outperformed PGGAN consistently

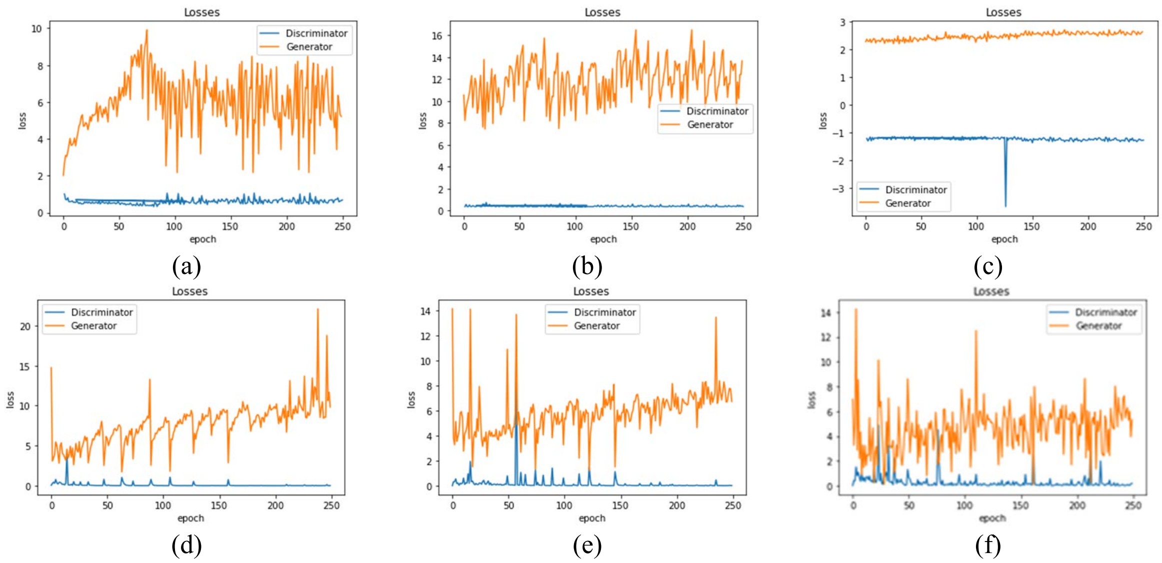

FIGURE 6. Training loss plots of PGGAN for (a) 8 × 8, (b) 16 × 16, (c) 32 × 32, (d) 64 × 64, (e) 128 × 128 and (f) 256 × 256 scale.

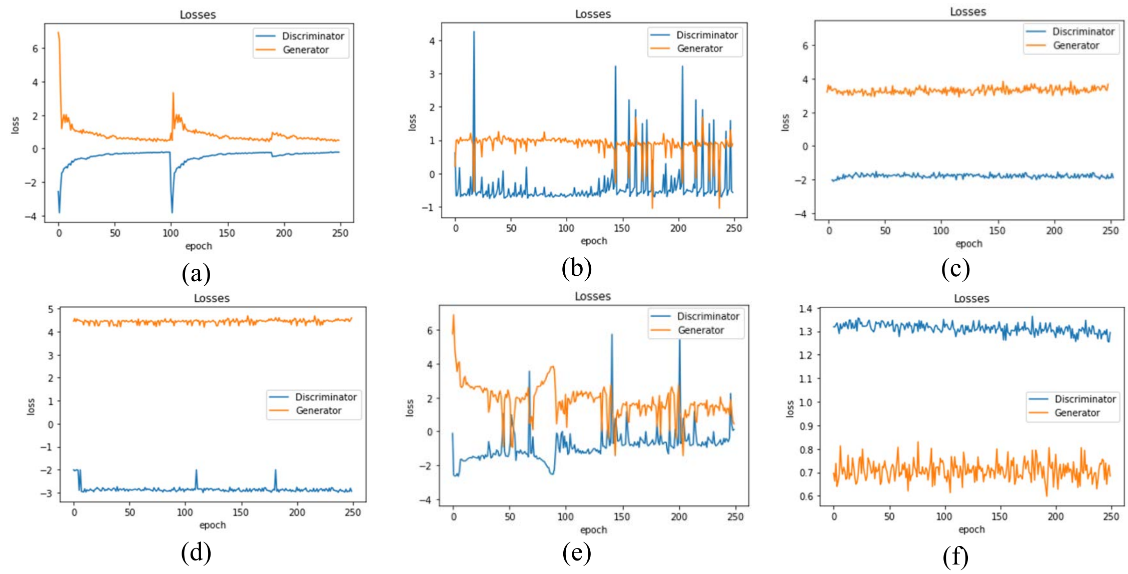

FIGURE 7. Training loss plots of HieGAN for (a) 8 × 8, (b) 16 × 16, (c) 32 × 32, (d) 64 × 64, (e) 128 × 128 and (f) 256 × 256 scale.

in both assessments by using MAE (256 × 256 - HieGAN: 0.00135, AEGAN: 0.00159, PGGAN: 0.00162, WGAN-GP: 0.00171) and WD (256 × 256 - HieGAN: 0.506, AEGAN: 0.573, PGGAN: 0.590, WGAN-GP: 0.612). We did not consider the unusual good WD scores in 16 × 16 as the size of synthetic image at that level still could not exhibit the knee structure. A series of real and synthetic knee images (produced by HieGAN) were exhibited in Fig. 8.

In Table 2 and 3, quality assessments of real and synthetic images were computed in 128 × 128 and 256 × 256 scales, respectively. The quality of synthetic knee image improved continuously from 128 × 128 until 256 × 256. The best AM score in 128 × 128 was 3.039 and improved until 2.419 in 256 × 256. Similar trend was observed in Mode score. The best score was recorded at 1.019 in 128 × 128 and improved until 1.383 in 256 × 256. HieGAN has demonstrated better capability in producing good quality synthetic image than other GANs in both scales.

IV. DISCUSSION

This is a novel work on knee image synthesis framework for GAN. The model performance of HieGAN was compared with other state-of-art i.e., WGAN-GP, PGGAN and AEGAN at different scales. We first assessed the data distribution difference between real and synthetic knee images by using MAE and WD. Then, we validated the quality and variation

TABLE 1. Probability distribution difference between real and synthetic knee images using distance measurements.

	Method	Scale					
		8 × 8	16 × 16	32 × 32	64 × 64	128 × 128	256 × 256
MAE	WGAN-GP	0.0489	0.0274	0.0193	0.00597	0.00421	0.00171
	PGGAN	0.0410	0.0240	0.0189	0.00451	0.00341	0.00162
	AEGAN	0.0406	0.0235	0.0186	0.00437	0.00339	0.00159
	HieGAN	0.0323	0.0150	0.0097	0.00310	0.00234	**0.00135**
WD	WGAN-GP	0.724	0.438	0.877	0.965	0.949	0.612
	PGGAN	0.626	0.351	0.897	0.987	0.908	0.590
	AEGAN	0.614	0.345	0.865	0.966	0.895	0.573
	HieGAN	0.546	0.328	0.857	0.996	0.865	**0.506**

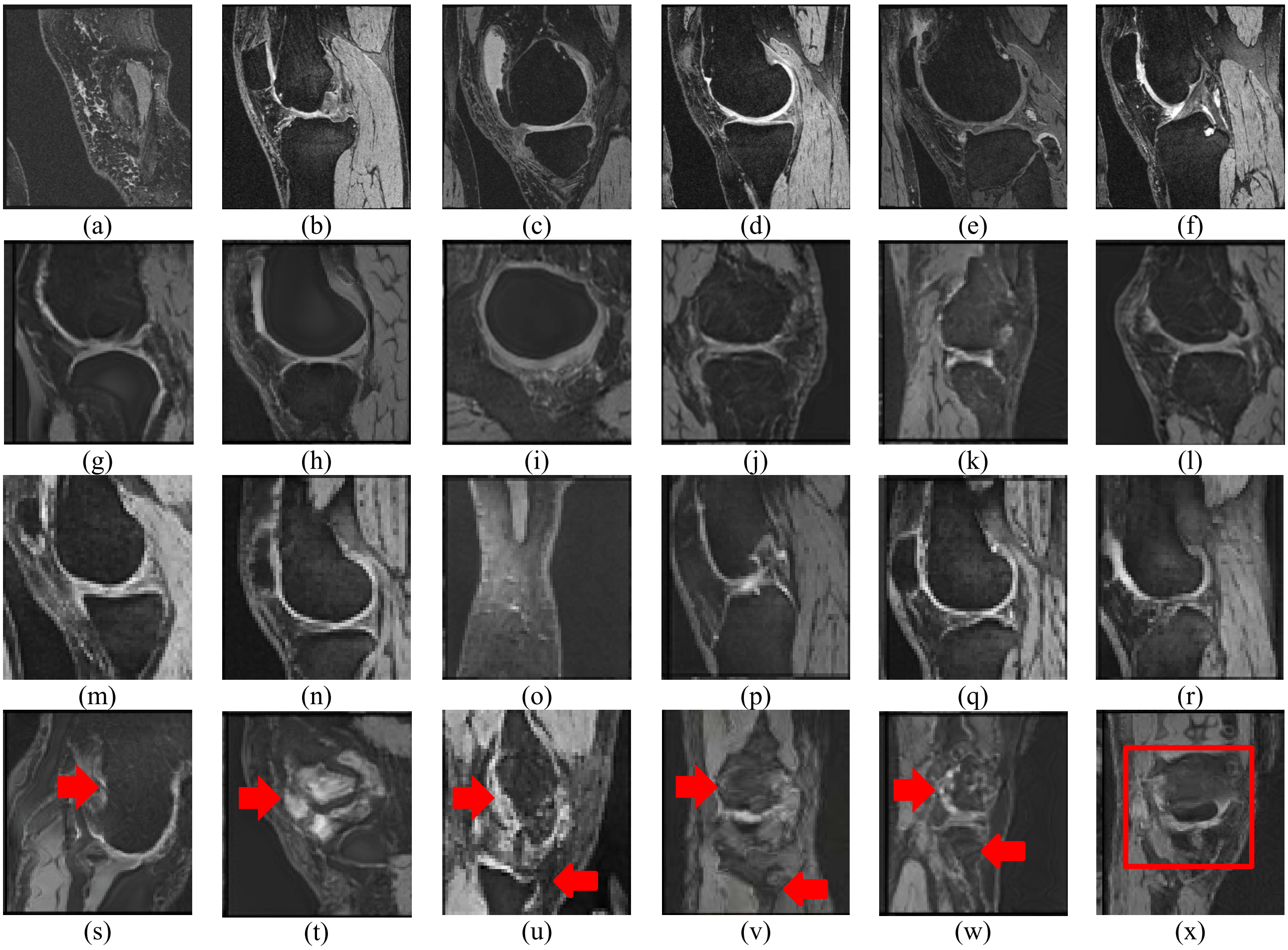

FIGURE 8. Comparison of real MR image of knee (a) to (f) against synthetic knee image at 128 _ 128 scale (g) to (l) and 256 _ 256 scale (m) to (r). Failure cases were indicated by red arrows in PGGAN (s) to (w) and by red box in HieGAN (x).

of synthetic knee image of 128 × 128 and 256 × 256 scales at different iterations by using AM and Mode score. Knee image synthesis is a challenging task because of its complex structure and varying anatomical geometry [30], [31]. Equipped with a novel normalization technique and attention configuration, we have shown that the proposed framework has successfully produced realistic synthetic knee images. In the following, we describe the key lessons obtained from this work.

First, synthetic knee images are useful in segmentation tasks. One potential application includes the diversification of real training data with synthetic data to improve the robustness of deep learning segmentation model. According to Russ *et al.* (2019), three training configurations i.e. real

TABLE 2. Assessment of image quality between real and synthetic knee images at 128 × 128 scale.

	Method	Iterations					
		38,000	76,000	114,000	152,000	190,000	228,000
AM	WGAN-GP	3.554	3.459	3.322	3.292	3.282	3.286
	PGGAN	3.219	3.199	3.135	3.205	3.225	3.204
	AEGAN	3.153	3.094	3.095	3.065	3.160	3.141
	HieGAN	3.126	**3.039**	3.125	3.094	3.073	3.067
Mode Score	WGAN-GP	0.697	0.763	0.830	0.784	0.815	0.823
	PGGAN	0.785	0.842	0.867	0.809	0.961	0.964
	AEGAN	0.796	0.866	0.829	0.820	0.853	0.912
	HieGAN	0.805	**1.019**	0.803	0.827	0.775	0.823

TABLE 3. Assessment of image quality between real and synthetic knee images at 256 × 256 scale.

	Method	Iterations					
		38,000	76,000	114,000	152,000	190,000	228,000
AM	WGAN-GP	3.674	3.615	3.474	3.828	3.458	2.861
	PGGAN	3.319	3.321	3.289	3.351	3.265	2.611
	AEGAN	3.254	3.216	3.185	3.176	3.089	2.547
	HieGAN	3.179	3.009	3.074	3.025	3.010	**2.419**
Mode Score	WGAN-GP	0.809	0.821	0.883	0.865	0.906	0.929
	PGGAN	0.883	0.982	0.968	0.891	0.984	1.009
	AEGAN	0.896	0.994	1.005	0.987	1.019	1.041
	HieGAN	1.011	1.026	0.990	1.126	0.995	**1.383**

data only, synthetic data only and real-synthetic data combination, were used to augment the training data of u-net segmentation model. Real-synthetic training data configuration had reported superior performance compared to previous two configurations [32]. Given the rising number of research works in knee segmentation using deep learning models [33], further investigations based on their findings will benefit future knee segmentation models.

Second, optimization of GAN training remains an active research topic despite its appealing potentials. In fact, conventional GANs are infamous for demonstrating mode collapse and vanishing gradient issues during the training process. Selection of normalization techniques has profound effect on the quality of knee image synthesis. For instance, the use of standalone spectral and pixelwise normalization have produced low quality knee images with blur background or inferior contrast, which cannot be adopted into subsequent deep learning segmentation models. Despite PGGAN has reported more stable training performance attributed to its progressive training nature, the model still suffers from training instability. PGGAN does not converge smoothly in some several scales. Meanwhile, original WGAN [16] was proposed to bring stability into GAN's training process but the outcome is highly dependent on the tuning of hyperparameter. As a result, WGAN might generate synthetic image with inferior quality while the training still fails to converge.

In recognition of the limitation, WGAN-GP uses gradient penalty to enforce the Lipschitz constraint. Nevertheless, WGAN-GP still lacks the capability to produce synthetic knee image with good quality. On the other hand, HieGAN manages to converge successfully at 256 × 256 scale after improvement was deployed on the model training stability. In addition, huge computational cost incurred by GANs is another topic of research. StyleGAN [34] is an extension of PGGAN. It has generated high-resolution attributes in natural images. Recently, it is extended to synthesize CT and MR images [35]. Unfortunately, implementation of StyleGAN requires extremely huge computational resources [36]. Thus, it is infeasible for common medical image synthesis applications.

AEGAN aims to tackle the problem of low-quality synthetic image by encoding the global structure features and extract salient image details. Based on the results, the model shows promising results in knee image synthesis. However, it is noteworthy that AEGAN consumes high computational cost at the expense of capturing fine details. In HieGAN, attention was employed as an alternative to capture salient features. The attention layer has successfully guided the discriminator to pay more attention to different features of knee images in order to compel the generator to produce realistic images without imposing extra computational burden to the training. Our quantitative findings suggested that the images have attained high degree of realism especially at 256 × 256 scale. Specifically, the overall image brightness is preserved, the boundary of cartilage-bone interface is well-preserved, the contrast between bone, cartilage and background is distinct, and the anatomical shape and size of cartilage and bone are conserved.

We have detected failure cases from samples generated by PGGAN. As such, PGGAN have generated seriously deformed knee structure wherein the structure of femur and tibia have been altered. Moreover, the boundary between femur and surrounding musculoskeletal tissues is overly diffused in several samples. The failure samples with serious deformation could potentially mislead the learning of deep learning models. Nonetheless, we also observed minor irregularity in one sample produced by HieGAN. The proposed model failed to distinguish between shrinking femur and tibia from the background musculoskeletal tissues. The boundary between knee bones and background is considered blur. These failure cases provide valuable insights for us to improve the model in the future.

At current stage, the study has some limitations. We do not assess the pathological feature of synthetic knee image. Some radiological features of knee osteoarthritis (OA) include osteophytes, bone marrow lesions and subchondral bone cysts are not taken into consideration. Besides, we decided the image generation until 256×256 scale in order to better understand the balanced results under the consideration of existing GPU capacity, salient feature recognition and acceptable medical image resolution scales. These factors are common among researchers in deep learning for medical image analysis field to build a sustainable medical image synthesis framework for GAN.

V. CONCLUSION

Research interest in GAN model development is growing among medical image analysis community along with the advances in GPU technology. Both quantitative and qualitative results showed that the HieGAN outperformed state-of-art PGGAN. In future, generation of synthetic knee image at higher resolution of 512×512 and 1024×1024 will be attempted by installing more powerful GPUs. In future work, Visual Turing Test will be conducted to investigate the capability of HieGAN in generating pathological features from knee images in collaboration with medical image experts. Then, the synthetic data will be tested along with real data to perform deep learning segmentation in an attempt to mitigate the curse of lack of training data faced by supervised deep learning models.

REFERENCES

[1] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552, doi: 10.1016/j.media.2019.101552.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," presented at the Adv. Neural Inf. Process. Syst., Montreal, QC, Canada, 2014.

[3] N. K. Singh and K. Raza, "Medical image generation using generative adversarial networks: A review," in *Health Informatics: A Computational Perspective in Healthcare*, R. Patgiri, A. Biswas, and P. Roy, Eds. Singapore: Springer, 2021, pp. 77–96.

[4] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341, doi: 10.1109/CVPR.2019.00244.

[5] T. Zhang, H. Fu, Y. Zhao, J. Cheng, M. Guo, Z. Gu, B. Yang, Y. Xiao, S. Gao, and J. Liu, "SkrGAN: Sketching-rendering unconditional generative adversarial networks for medical image synthesis," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (Lecture Notes in Computer Science), vol. 11767. Shenzhen, China: Springer, 2019, pp. 777–785.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," presented at the 4th Int. Conf. Learn. Represent., San Juan, PR, USA, May 2016.

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[8] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 240–244, doi: 10.1109/ISBI.2018.8363564.

[9] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 289–293, doi: 10.1109/ISBI.2018.8363576.

[10] A. Beers, J. Brown, K. Chang, J. P. Campbell, S. Ostmo, M. F. Chiang, and J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," 2018, *arXiv:1805.03144*.

[11] R. Togo, T. Ogawa, and M. Haseyama, "Synthetic gastritis image generation via loss function-based conditional PGGAN," *IEEE Access*, vol. 7, pp. 87448–87457, 2019, doi: 10.1109/ACCESS.2019.2925863.

[12] G.-P. Diller, J. Vahle, R. Radke, M. L. B. Vidal, A. J. Fischer, U. M. M. Bauer, S. Sarikouch, F. Berger, P. Beerbaum, H. Baumgartner, and S. Orwat, "Utility of deep learning networks for the generation of artificial cardiac magnetic resonance images in congenital heart disease," *BMC Med. Imag.*, vol. 20, no. 1, p. 113, Dec. 2020, doi: 10.1186/s12880-020-00511-1.

[13] H. Y. Park, H.-J. Bae, G.-S. Hong, M. Kim, J. Yun, S. Park, W. J. Chung, and N. Kim, "Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: Visual Turing test," *JMIR Med. Informat.*, vol. 9, no. 3, Mar. 2021, Art. no. e23328, doi: 10.2196/23328.

[14] D. Korkinof, H. Harvey, A. Heindl, E. Karpati, G. Williams, T. Rijken, P. Kecskemethy, and B. Glocker, "Perceived realism of high-resolution generative adversarial network–derived synthetic mammograms," *Radiol., Artif. Intell.*, vol. 3, no. 2, Mar. 2021, Art. no. e190181, doi: 10.1148/ryai.2020190181.

[15] B. Segal, D. M. Rubin, G. Rubin, and A. Pantanowitz, "Evaluating the clinical realism of synthetic chest X-rays generated using progressively growing GANs," *Social Netw. Comput. Sci.*, vol. 2, no. 4, p. 321, Jul. 2021, doi: 10.1007/s42979-021-00720-7.

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," presented at the 34th Int. Conf. Mach. Learn., 2017. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[17] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2726–2737, Nov. 2019, doi: 10.1109/TMM.2019.2908352.

[18] H.-S. Gan, K. A. Sayuti, N. H. Harun, and A. H. A. Karim, "Flexible non cartilage seeds for osteoarthritic magnetic resoance image of knee: Data from the osteoarthritis initiative," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2016, pp. 748–751.

[19] H.-S. Gan, T.-S. Tan, K. A. Sayuti, A. H. A. Karim, and M. R. A. Kadir, "Multilabel graph based approach for knee cartilage segmentation: Data from the osteoarthritis initiative," in *Proc. IEEE Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2014, pp. 210–213, doi: 10.1109/IECBES.2014.7047487.

[20] Z. Zhang, M. Li, and J. Yu, "D2PGGAN: Two discriminators used in progressive growing of GANS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3177–3181, doi: 10.1109/ICASSP.2019.8683262.

[21] X. Gong, S. Chang, Y. Jiang, and Z. Wang, "AutoGAN: Neural architecture search for generative adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Long Beach, CA, USA, Oct. 2019, pp. 3224–3234.

[22] F. Eckstein, "Double echo steady state magnetic resonance imaging of knee articular cartilage at 3 tesla: A pilot study for the osteoarthritis initiative," *Ann. Rheumatic Diseases*, vol. 65, no. 4, pp. 433–441, Apr. 2006, doi: 10.1136/ard.2005.039370.

[23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," presented at the 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," presented at the 6th Int. Conf. Learn. Represent., Vancouver, BC, Canada, 2018.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 18–23.

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," presented at the 30th Int. Conf. Neural Inf. Process. Syst., Barcelona, Spain, 2016.

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," presented at the Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.

[28] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," presented at the 5th Int. Conf. Learn. Represent., Toulon, France, 2017.

[29] Z. Zhou, H. Cai, S. Rong, Y. Song, K. Ren, W. Zhang, Y. Yu, and J. Wang, "Activation maximization generative adversarial nets," presented at the 6th Int. Conf. Learn. Represent., Vancouver, BC, Canada, 2018.

[30] G. Hong-Seng, K. A. Sayuti, and A. H. A. Karim, "Investigation of random walks knee cartilage segmentation model using inter-observer reproducibility: Data from the osteoarthritis initiative," *Bio-Med. Mater. Eng.*, vol. 28, no. 2, pp. 75–85, Mar. 2017, doi: 10.3233/BME-171658.

[31] H. S. Gan and K. A. Sayuti, "Comparison of improved semi-automated segmentation technique with manual segmentation: Data from the osteoarthritis initiative," *Amer. J. Appl. Sci.*, vol. 13, no. 11, pp. 1068–1075, Nov. 2016.

[32] T. Russ, S. Goerttler, A.-K. Schnurr, D. F. Bauer, S. Hatamikia, L. R. Schad, F. G. Zöllner, and K. Chung, "Synthesis of CT images from digital body phantoms using CycleGAN," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 10, pp. 1741–1750, Oct. 2019, doi: 10.1007/s11548-019-02042-9.

[33] H.-S. Gan, M. H. Ramlee, A. A. Wahab, Y.-S. Lee, and A. Shimizu, "From classical to deep learning: Review on cartilage and bone segmentation techniques in knee osteoarthritis research," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 2445–2494, Apr. 2021, doi: 10.1007/s10462-020-09924-4.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[35] L. Fetty, M. Bylund, P. Kuess, G. Heilemann, T. Nyholm, D. Georg, and T. Löfstedt, "Latent space manipulation for high-resolution medical image synthesis via the StyleGAN," *Zeitschrift Medizinische Physik*, vol. 30, no. 4, pp. 305–314, Nov. 2020, doi: 10.1016/j.zemedi.2020.05.001.

[36] A. H. Bermano, R. Gal, Y. Alaluf, R. Mokady, Y. Nitzan, O. Tov, O. Patashnik, and D. Cohen-Or, "State-of-the-art in the architecture, methods and applications of StyleGAN," 2022, *arXiv:2202.14020*.

HONG-SENG GAN (Senior Member, IEEE) received the B.Eng. degree in biomedical and the Ph.D. degree in biomedical engineering from Universiti Teknologi Malaysia, in 2012 and 2016, respectively. His research interests include medical image processing, artificial intelligence, machine learning, and computer vision. He is currently a Faculty Member of the Department of Data Science, Universiti Malaysia Kelantan.

MUHAMMAD HANIF RAMLEE received the B.Eng. degree in biomedical and the Ph.D. degree in biomedical engineering from Universiti Teknologi Malaysia, in 2012 and 2016, respectively. His research interests include biomechanics computation and stimulation. He is currently a Faculty Member of the School of Biomedical Engineering and Health Sciences, Universiti Teknologi Malaysia.

BANDER ALI SALEH AL-RIMY (Senior Member, IEEE) received the B.Eng. degree in computing from Sana'a University, in 2003, the M.Sc. degree in information technology from Open University Malaysia, in 2013, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia, in 2019. His research interests include computer networks and artificial intelligence. He is currently a Faculty Member of the School of Computing, Universiti Teknologi Malaysia.

YENG-SENG LEE received the B.Eng. and Ph.D. degrees in communication engineering from Universiti Malaysia Perlis, in 2012 and 2016, respectively. His research interests include signal processing, electromagnetic and dielectric material characterization. He is currently a Faculty Member of the Faculty of Electronics Engineering Technology, Universiti Malaysia Perlis.

PRAYOOT AKKARAEKTHALIN received the B.Eng. and M.Eng. degrees in electrical engineering from the King Mongkut's Universsity of Technology North Bangkok (KMUTNB), Thailand, in 1986 and 1990, respectively, and the Ph.D. degree from the University of Delware, Newark, USA, in 1998. His research interests include signal processing and wideband and multiband antennas. He is currently a Faculty Member of the Faculty of Engineering, KMUTNB.