GANai: Standardizing CT Images using Generative Adversarial Network with Alternative Improvement

Gongbo Liang[1,2], Sajjad Fouladvand[1,2], Jie Zhang[3], Michael A. Brooks[3], Nathan Jacobs[2], Jin Chen[1,2,4]

1 Institute for Biomedical Informatics, University of Kentucky, USA, Lexington, KY, USA
2 Department of Computer Science, University of Kentucky, Lexington, KY, USA
3 Department of Radiology, University of Kentucky, Lexington, KY, USA
4 Department of Internal Medicine, University of Kentucky, Lexington, KY, USA

***Abstract*—Computed tomography (CT) is a widely-used diagnostic image modality routinely used for assessing anatomical tissue characteristics. However, non-standardized imaging protocols are commonplace, which poses a fundamental challenge in large-scale cross-center CT image analysis. One approach to address the problem is to standardize CT images using generative adversarial network models (GAN). GAN learns the data distribution of training images and generate synthesized images under the same distribution. However, existing GAN models are not directly applicable to this task mainly due to the lack of constraints on the mode of data to generate. Furthermore, they treat every image equally, but in real applications, some images are more difficult to standardize than the others. All these may lead to the lack-of-detail problem in CT image synthesis. We present a new GAN model called GANai to mitigate the differences in radiomic features across CT images captured using non-standard imaging protocols. Given source images, GANai composes new images by specifying a high-level goal that the image features of the synthesized images should be similar to those of the standard images. GANai introduces an alternative improvement training strategy to alternatively and steadily improve model performance. The new training strategy enables a series of technical improvements, including phase-specific loss functions, phase-specific training data, and the adoption of ensemble learning, leading to better model performance. The experimental results show that GANai is significantly better than the existing state-of-the-art image synthesis algorithms on CT image standardization. Also, it significantly improves the efficiency and stability of GAN model training.**

Index Terms–**computed tomography, image synthesis, generative adversarial network, alternative training**

I. Introduction

Computed tomography (CT) is one of the most popular diagnostic image modalities routinely used for assessing anatomical tissue characteristics for disease management [1], [2], [3], [4], [5]. CT scanners provide the flexibility of customizing acquisition and image reconstruction protocols to meet an individual's clinical needs [6], [7]. However, CT acquisition parameter customization is a double-edged sword [8]. While it enables physicians to capture critical image features towards personalized healthcare, it forms a barrier to analyzing CT images in a large scale, in that capturing CT images with non-standardized imaging protocols may result in inconsistent radiomic features [9], [10]. As was revealed in a recent study, both intra-CT (by changing CT acquisition parameters) and inter-CT (by comparing different scanners with the same acquisition parameters) tests have demonstrated low reproducibility regarding radiomic features, such as intensity, shape, and texture, for CT imaging [11], [12]. In the example shown in Figure 1, each lung tumor was acquired twice using two different reconstruction kernels (Bl64 and Br40, Siemens Healthineers, Erlangen, Germany). The figure demonstrates that the appearances (as well as the radiomic features) of the same tumor can be strongly affected by the selection of CT acquisition parameters.

To overcome the barriers that prevent the use of CT images in large-scale radiomic studies, algorithms have been developed aiming to integrate and standardize CT images from multiple sources. Image synthesis is a class of algorithms that generate synthesized images from source images, which satisfy the condition that the feature-based distributions of the synthesized images are similar to that of target images [13]. Mathematically, given source image x, an image synthesis algorithm composes a synthesized image x' by specifying a *high-level* goal that the image features of x' are significantly more similar to that of the target image y than the source image x. Image synthesis algorithms have been widely used in image conversion and natural language processing, such as the synthesis of images from text descriptions [14]. Note that image synthesis is different from image conversion (such as to convert an MRI image to a CT image), which requests an exact pixel-to-pixel match between the synthesized images and the target images [15].

Image synthesis algorithms can be roughly classified into two groups, i.e., traditional image processing algorithms and deep learning-based algorithms. In the first group, the histogram matching-based algorithm has been widely used [16], [17], [18], [19]. In general, it synthesizes images by mapping the histogram of source images to that of target images. However, finding the mapping function requires the presence of the target images, which are often missing or are not well defined in practice. In the second group, generative adversarial network models (GAN), a class of deep learning algorithms, can learn the data distribution of training data and generate synthesized examples which fall under the same distribution of the training [20]. In particular, the conditional generative adversarial network (cGAN), a special kind of GANs, learns the conditional distribution of the source image x given the

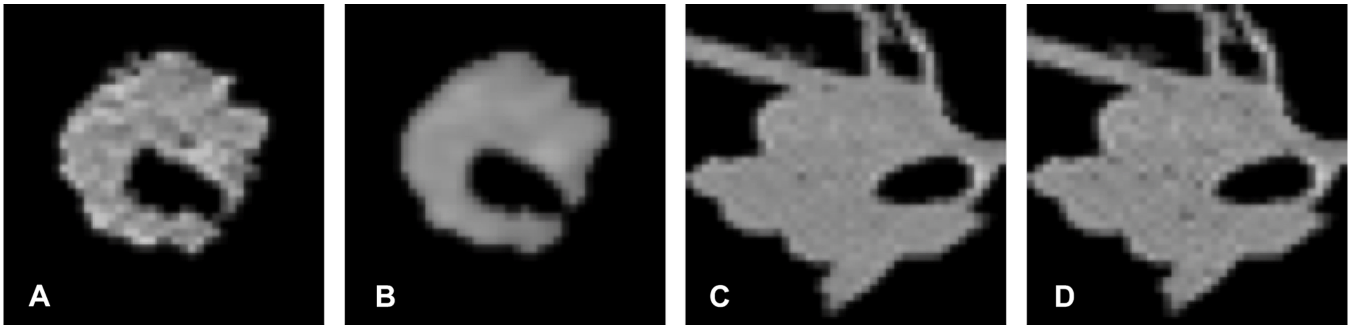

Fig. 1: Lung tumors acquired using two kernels have shown significantly different appearances as well as radiomic features. (A) Lung tumor 1 acquired with kernel Bl64. (B) Lung tumor 1 acquired with kernel Br40. (C) Lung tumor 2 acquired with kernel Bl64. (B) Lung tumor 2 acquired with kernel Br40.

target image y and then performs image transference from one domain to another [21], [22]. However, GAN models (include cGAN) are not directly applicable to our task mainly due to three limitations: 1) GAN models do not contain any constraints to control what modes of data it shall generate; 2) the synthesized images are not guaranteed to be similar to the target images (Figure S1); 3) GAN models treat every image in training equally, but in real applications, some images are more difficult to synthesize than the others (Figure S2). All these limit the functionality of GAN models and may lead to the lack-of-detail problem in image synthesis.

To address the computational challenges in medical image synthesis, where great image details have to be maintained, we propose a novel deep learning framework called "Generative Adversarial Network with Alternative Improvement (GANai)". GANai has a similar architecture as cGAN, but its training process is significantly different. Specifically, GANai introduces an alternative improvement training strategy to alternatively train its deep learning components and steadily improve the whole model performance. The adoption of the new training strategy enables a series of technical improvements, including phase-specific loss functions, component-dedicate training data, adoption of ensemble learning, and so on, leading to a significant improvement on model performance.

While GANai can be deployed in many applications, we adopted and evaluated GANai in mitigating the differences in radiomic features due to using non-standardized CT imaging protocols. The experimental results show that GANai is significantly better than the state-of-the-art image synthesis algorithms, such as cGAN and histogram matching, on all the image acquisition parameters that we have tested. In summary, GANai has the following computational advantages:

1) GANai introduces an alternative improvement training strategy to alternatively and steadily improve model performance.
2) GANai adopts a new phase-specific loss function that allows the discriminator and the generator to collaborate rather than competing with each other.
3) GANai improves model training effectiveness by training the discriminator and the generator using specified training images.
4) GANai adopts ensemble learning to significantly improve the stability of GAN model training .

II. Background

Radiomics is an emerging science to extract and use comprehensive radiomic features from a large volume of medical images for the quantification of overall tumor spatial complexity and the identification of tumor subregions that drive disease transformation, progression, and drug resistance [23], [24], [25], [26]. However, due to the use of non-standardized imaging protocols, variations in acquisition and image reconstruction parameters may cause inconsistency in radiomic features extracted from images, which poses a barrier to the practice of radiomics in large-scale [10], [24], [25].

A. CT Image Acquisition Parameters

In modern CT imaging, there are a large number of imaging protocols, and using non-standardized imaging protocols is common [6]. The CT image acquisition parameters includ kV (the x-tube voltage), mAs (the product of x-ray tube current and exposure time), collimation, pitch, reconstruction kernel, field-of-view, and slice thickness [27], [28]. In routine clinical practice, certain parameters are often adjusted to meet the diagnostic needs, i.e., to obtain satisfactory image quality while maintaining low radiation dose to patients. Changing acquisition parameters may significantly affect the resulting images (Figure 1). For example, adjusting kV will change CT numbers (the pixel values of a CT image), changing mAs will affect image noise rate, and the selection of reconstruction algorithms will result in different image texture features.

B. Histogram Matching

Histogram matching (or called histogram specification) is a widely-used image synthesis tool. It uses the intensity histogram to represent images and then transforms a source image to a target image by matching their intensity histograms [16], [17], [18], [19]. While histograms can represent the density of intensity in the whole image, the major drawback is the loss of location information. A variation of histogram matching is to divide a source image into multiple patches and to apply histogram matching on each patch, expecting that such patch-based representation may lead to location-specific image synthesis. However, patch-based histogram matching may introduce artifacts, esp. on the edges of patches. It is also sensitive to the selection of matching parameters such as the number of bins of a histogram (Figure S3).

C. Generative Adversarial Networks

Recently, deep learning has shown remarkable performance in various medical informatics tasks. For example, it has surpassed the human experts' performance on skin cancer classification by only looking at the dermoscopic images [29].

The generative adversarial network (GAN) is a kind of deep learning models that learns the data distribution of training images and generate synthesized images under the same distribution [20], [30], [31]. A GAN model usually has two components, i.e., the discriminator (D) and the generator (G), where G generates synthesized data from random noise, and D learns a data distribution from the training data and determines whether the synthesized data generated by G fall into the distribution. The goal of G is to generate synthesized data which are good enough to fool D, while D always aims to discriminate the synthesized data and the real data.

The conditional generative adversarial network (cGAN) is a kind of GAN models that learns the *conditional* distribution of the training data and generates synthesized data under the same condition [21], [32], [33]. Among cGAN models, the Image-to-Image model performs the image-to-image transference from one domain to another concerning the given condition, and it has become a widely recognized conditional image synthesis model [22]. Note that the images synthesized by cGANs are not necessarily similar to the target images, although they look "real", meaning having similar semantic meanings as the target images (see Figure S1, S2). However, in medical applications, it is important to maintain authenticity in the synthesized CT images. Specifically, it is expected to generate images with the distribution of radiomic features significantly similar to that of the target images.

While GAN models are advanced in image synthesis [22], [14], image inpainting [34], semantic segmentation [35], etc., GAN models are suboptimal regarding training efficiency and stability. To address the GAN training problem, several ensemble learning-based strategies have been applied to improve model training: 1) to train multiple GANs in parallel using a random initialization of model parameters, and then to randomly choose one of the GANs to generate the synthesized data [36]; 2) to train multiple Ds and requires the G to fool a group of Ds [37]; and 3) to select training data using boosting and to train a cascade of GANs in sequence. It has been shown that the performance of GANs can be significantly improved by using ensemble learning [37].

III. METHOD

To extend the adversarial learning into the medical image domain and to address the aforementioned challenges, we propose Generative Adversarial Network with Alternative Improvement (GANai).

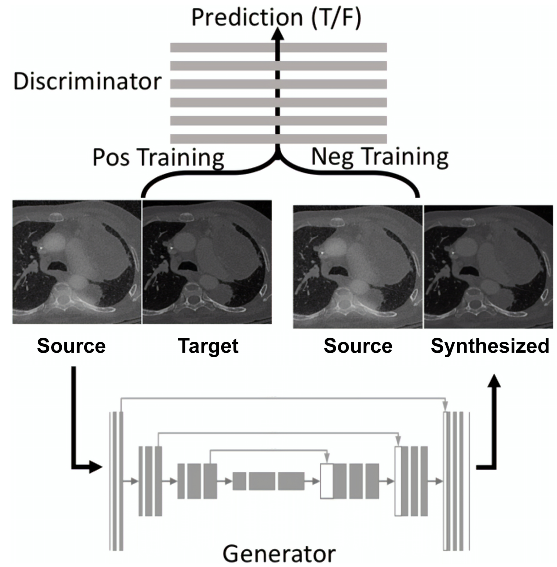

Fig. 2: Architecture of GANai. Given a source image, the generator G synthesizes a new image to fool the discriminator D, while D aims to distinguish the synthesized image and the target image.

A. Architecture

GANai consists of two components, i.e., the generator (G) and the discriminator (D), where G is a U-Net with fifteen hidden layers and D is a multilayer perceptron model with six fully connected layers [38]. The architecture of GANai is similar to the cGAN models, shown in Figure 2 [22]. The inputs of D of GANai are image pairs (x, y) and (x, x'), where (x, y) denotes the real pair (positive training), and (x, x') denotes the fake pair (negative training). The goal of D is to distinguish the real pairs from the fake pairs. Given the feedback from D, G learns the mapping from X to Y and generates a synthesized image x' for any given source image x ($x \in X$) in Y's domain. In contrast to D, G aims to synthesize images that can fool D. If D can distinguish most of the fake pairs from the real pairs, the performance of G needs to be further improved. Otherwise, we conclude that the generative results of G are good enough for the current D.

B. Alternative Improvement

In traditional GAN models, D and G are trained synchronously (D and G trained together) or asynchronously (several batches of D-training followed by several batches of G-training), based on the assumption that both D and G can be gradually improved together. In practice, however, if D is not well trained to capture the intrinsic features to separate a real and a fake image, G can easily fool D. Similarly, if G is not well "challenged" by D, its model performance is not guaranteed to be improved.

We introduce the alternative training approach for GANs (Figure 3). As the name suggested, GANai has two alternate training phases, i.e., the discriminator training (D-training) and the generator training (G-training). In each training phase, we focus on optimizing one of the components while freezing the other. A training phase will stop if the current component

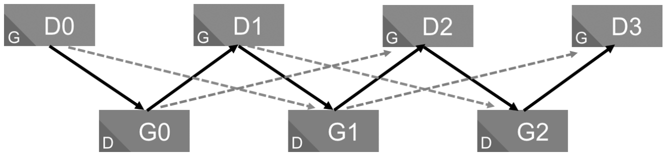

Fig. 3: In each training phase of GANai, D (or G) is trained while the other component is frozen. The name of a block (such as D_0 or G_1) indicates the component in training, and the letter located at the bottom left corner indicates the component that is frozen. The alternative training (solid line) ensures high performance while the ensemble approach (dotted line) improves the training stability.

is well trained or the training step exceeds an upper bound (see Section V-B for more details). After that, we switch to the other training phase (Figure 3 solid lines). The alternative training strategy enables a series of technical improvements, including phase-specific loss functions, phase-specific training data, and the adoption of ensemble learning, which will be introduced in the following subsections.

C. Loss Functions

The alternative training of GANai may boost model performance by preventing each component being too strong or too weak. In the literature, strategies have been presented to freeze part of a GAN when the GAN components are imbalanced [39]. However, it is difficult to decide when to freeze/unfreeze a component of GAN. To address this issue, we redesigned the loss functions.

In the D-training phase, G is frozen so that D learns the differences between the synthesized images and the target images and discriminates the synthesized images. Hence, the loss function of D is the same discriminator loss of cGAN [21]:

$$\begin{aligned}\mathbb{L}_{Phase\_D}(D) =& \mathbb{E}_{x,y\sim P_{data}(x,y)}[-\log D(x,y)]+\\ & \mathbb{E}_{x\sim p_x, z\sim p_z(z)}[-\log(1-D(x,G(x,z)))]\end{aligned} \tag{1}$$

where x is the source image; y is the target image; $G(x,z)$ is the synthesized image generated by G, which maps the source image x and a random noise vector z to y; $D(x,y)$ is the prediction result of the real pair; and $D(x,G(x,z))$ is the prediction result of the fake pair. For $D(x,y)$, the higher the prediction accuracy, the higher the value of $D(x,y)$.

In the G-training phase, D is frozen, and it evaluates the results of G. Since we expect G to fool D, the loss of D in the G-training phase is defined as:

$$\begin{aligned}\mathbb{L}_{Phase\_G}(D) =& \mathbb{E}_{x,y\sim P_{data}(x,y)}[-\log D(x,y)]+\\ & \mathbb{E}_{x\sim p_x, z\sim p_z(z)}[-\log(D(x,G(x,z)))]\end{aligned} \tag{2}$$

Finally, by integrating Eq 1 and Eq 2, the loss function of D in GANai is defined as:

$$\begin{aligned}\mathbb{L}(D) =& \mathbb{E}_{x,y\sim P_{data}(x,y)}[-\log D(x,y)]+\\ & \left(\mathbb{E}_{x\sim p_x, z\sim p_z(z)}[-\log D(x,G(x,z))]\right)^{\alpha}+\\ & \left(\mathbb{E}_{x\sim p_x, z\sim p_z(z)}[-\log(1-D(x,G(x,z)))]\right)^{1-\alpha}\end{aligned} \tag{3}$$

where parameter $\alpha = 1$ if GANai is in the G-training phase and $\alpha = 0$ in the D-training phase.

The loss function of G is the same as Isola et al. [22]. Also, we adopt the $L1$ loss as the regularization factor.

$$\begin{aligned}\mathbb{L}(G) =& \mathbb{E}_{x,G(x,z)}[-\log D(x,G(x,z))]+\\ & \beta\mathbb{E}_{G(x,z),y}[||y-G(x,z)||]\end{aligned} \tag{4}$$

where β is the weight of the regularization term.

To determine when to switch between the D-training phase and the G-training phase, the prediction accuracy on the fake image pairs ($D(x,x')$) is used. The value of $D(x,x')$ is computed at every training step and is compared with two thresholds. More specifically, if $D(x,x') \leq T_l$, GANai will switch from D-training to G-training. If $D(x,x') \geq T_h$, GANai will switch from G-training to D-training. T_l and T_h are the lower and upper thresholds of $D(x,x')$. To improve training stability, the least amount of steps (minibatches) of each training phase is also specified. Note that in GANai, the value of $D(x,x')$ increases and decreases, indicating that the performance of D and G is improved alternatively.

D. Training D and G with Dedicated Training Data

Since the components of GANai are trained separately, one idea is to increase model training efficiency by training G and D using different data. More specifically, the images that are potentially synthesizable can be used to accelerate the G-training, while the training of D can benefit from images that are difficult to synthesize.

We develop a procedure to select training data for D and G. First, a cGAN model is trained using all the training data [22]. Second, with the trained cGAN model, we synthesize a new image for every source image and compare every synthesized image with its corresponding target image using Kullback-Leibler divergence [40], normalized mutual information (N-MI) [41], and cosine similarity. Finally, the training data is split into two subsets based on z-score, i.e., $1/3$ of the source-target image pairs with the highest similarities between synthesized images and target images (called T_{easy}) and $1/3$ of the images with the lowest similarities (call T_{hard}). The new procedure allows us to train G using T_{easy} and train D using T_{hard} (see Section V-A for other training set selection strategies).

E. Improving Training Stability using Ensemble Learning

Due to the nature of the generative adversarial concept (i.e., open-ended competition between GAN components), it is not guaranteed that G or D will improve towards the same direction. For example, if the kth state of G fools the $(k-1)$th state of D, it still may be classified by the older $(k-2)$th state of D. Therefore, during the two-phase training of GANai, we improve the model stability by adopting the ensemble learning. Simply speaking, a D is required to discriminate multiple Gs and a G must fool multiple Ds.

Mathematically, the following criteria are specified in GANai: when training the kth G, the G must fool both the $(k-2)$th state and the $(k-1)$th state of D, and when training kth D, the D should discriminate both the $(k-2)$th state and the $(k-1)$th state of G. For an illustrative example, see the dot lines in Figure 3. These criteria can be further extended to incorporate more historical Ds or Gs or more sophisticated conditions. In the exception that GANai cannot identify such a D or G that satisfies the criteria after at most T_s steps (the maximum training step in each phase), it will roll back to the previous state, and re-train the current component.

IV. Experimental Results

A. Data

In total 2,448 chest CT image slices of lung cancer patients were collected using Siemens CT Somatom Force at the University of Kentucky Medical Center. For each patient, a CT image was constructed with each of the possible combinations of two image reconstruction parameters, i.e., slice thickness (0.5, 1, 1.5, 3mm) and reconstruction kernels (Bl57 and Bl64). With data augmentation, the training data has been extended to 14,958 image patch pairs. Among them, 7,479 assigned as T_{easy} and 7,479 assigned as T_{hard} using the procedure introduced in Section III-D. Each image pair contains a source image x and the target image y. See details of data augmentation in Section S1.A with examples in Figure S4.

The validation data contains 3,554 2.5D images, and multiple radiomic features were extracted for model validation. Specifically, we randomly cropped 2.5D images from the CT images that have not been used as training data, with their dimensions ranging from $5 \times 5 \times 5$ to $60 \times 60 \times 30$ pixels. When cropping the 2.5D validation images, we excluded areas with bone or air, since soft tissues are what physicians are most interested. See Section S1.B for more details.

Given a large number of CT imaging protocols, it is impractical to apply all of them. We selected two image reconstruction parameters (kernel and slice thickness) and used all the combinations for the model performance test. Also, we chose 1mm slice thickness and Bl64 kernel to be the standard imaging protocol, since it is widely used in the current lung cancer radiomic studies. Note the settings can be easily extended to incorporate more acquisition parameters or to use a different standardized imaging protocol.

B. Implementation Details of GANai

In GANai, G is a fifteen hidden layers U-Net [38], with the size between $128 \times 128 \times 64$ and $1 \times 1 \times 512$ (Figure S5). The input of G are 256×256 images, and the synthesized images have the same image size. D is implemented as a multilayer perceptron model with six fully connected layers with the size between $256 \times 256 \times 3$ and $30 \times 30 \times 1$ (Figure S6).

The training of GANai started with the D-training phase, and all the network weights were randomly initialized. We set the regularization term weight $\beta = 100$ to reduce the visual artifacts [22], and used $T_l = 0.05$ and $T_h = 0.95$ as the training phase switch thresholds, and $T_s = 10$ epochs as the maximum training step. Within each training phase, the model needed to be trained for at least five steps before switching to the other training phase. GANai was trained for 100 epochs with learning rate being 0.0002, momentum being 0.5.

GANai is deployed on Tensorflow [42] on a Linux computer server with eight Nvidia GTX 1080 GPU cards. It took 15 hours to train GANai from scratch using a single GPU card. Using the trained model, it took 0.2 seconds to generate a synthesized image (5 images per second).

Figure 4 shows the discriminator prediction results on all the fake pairs $D(x, x')$ in the first 150 steps of training. With the training of D, $D(x, x')$ decreases. When the value of $D(x, x')$ is below T_l (in our experiment, $T_l = 0.05$), GANai is switched to the G-training phase. In the G-training phase, $D(x, x')$ increases, since D is frozen and the performance of G keeps increasing. When the value of $D(x, x')$ is higher than T_h ($T_h = 0.95$), GANai is switched to the D-training phase.

The training and validation loss of D and G in the first 150 training steps are shown in Figure 5. Both the training and validation loss of D decreased in every training phase, which indicates the model performance of D and G was improved alternatively. In the D-training phase, if the performance of D is increased, the loss of D will reduce, since both $-\log(D(x, y))$ and $-\log(1 - D(x, x'))$ are both reduced (solid lines in Figure 5A). When switching from the D-training to the G-training phase, α in the loss function of D flips from 0 to 1, which immediately turns the loss of D from a small value to a high value (see the jumps located at phase turning points in Figure 5A). In the G-training phase, if the performance of G is increased, the performance of D will decrease, so the loss of D decreases (dotted lines in Figure 5A). Figure 5B shows the loss of G increases in D-training phase (due to the performance improvement of D) and decreases in G-training phase, since the performance of G is improved (See Section V-C).

C. Evaluation Metric

For performance evaluation, we compared GANai with cGAN [22] and the patch-based histogram matching (see details in supplementary section III). Instead of hiring human annotators, we adopt the radiomic features for performance evaluation [43], [44]. Specifically, two classes of radiomic features were used for model performance evaluation, i.e., 2.5D texture features (i.e., gray-level co-occurrence matrix) and 2.5D intensity histogram based features. In total, eight radiomic features were adopted for performance evaluation (see Section S2 for details).

Per every radiomic feature to test, we compared each synthesized image and its target image, and computed the absolute error and relative error using the following equations:

$$abs\_err(feature_k, m) = \frac{|feature(synthesized, k, m) - feature(target, k)|}{feature(target, k)} \quad (5)$$

where $feature_k$ is the kth radiomic feature, m is either GANai or a image synthesis model to compare.

$$rel\_err(feature_k, m_1, m_2) = \frac{abs\_err(feature_k, m_1) - abs\_err(feature_k, m_2)}{error(feature_k, m_1)}. \quad (6)$$

where m_1 and m_2 are two different image synthesis models. For the relative error, a positive value indicates that m_2 has smaller error than m_1, vice versa.

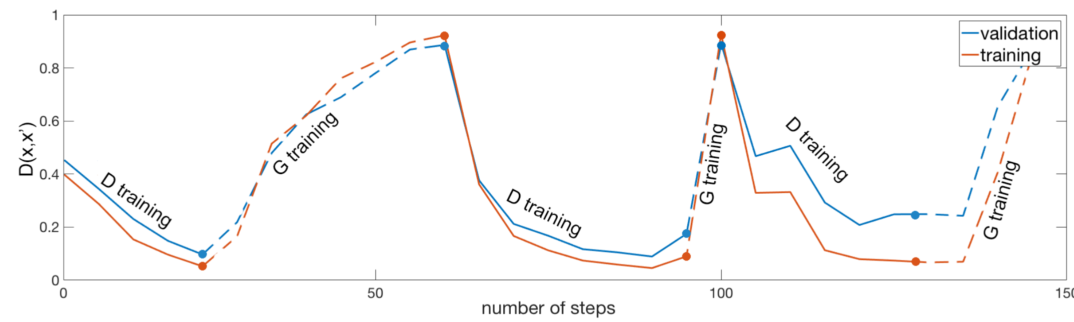

Fig. 4: The prediction results of D on the fake image pairs (x, x') in the first 150 steps of the alternative training. For $D(x, x')$, the higher the prediction accuracy, the lower the value ($D(x, x') \in [0, 1]$).

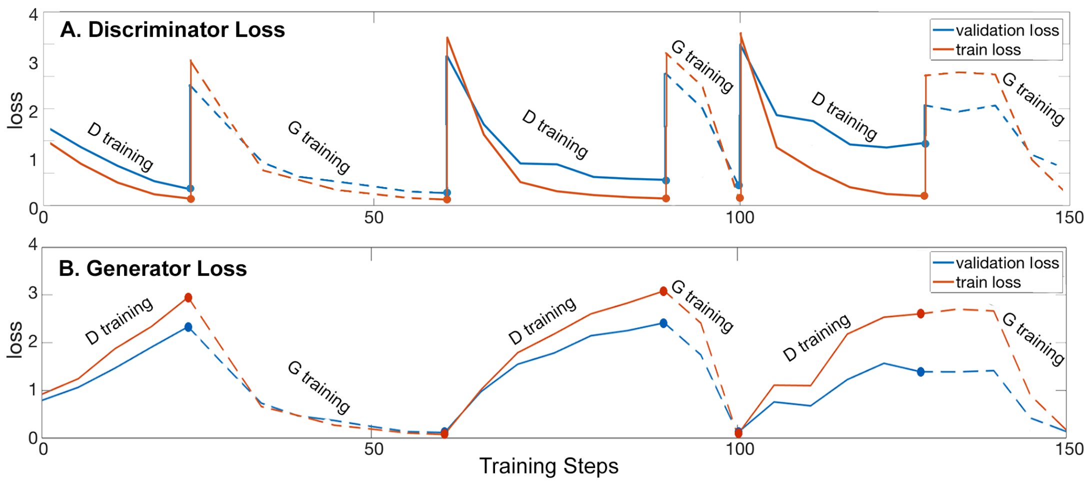

Fig. 5: The training loss and the validation loss of D and G in GANai in first 150 steps of training. The solid lines indicate the loss of D in the D-training phase. The dotted lines indicate the loss of D in the G-training phase. The solid points indicate the time when GANai switches between the D-training phase and the G-training phase.

Model stability is evaluated using the cumulative sum control chart (CUSUM) [45]. CUSUM is a sequential analysis model typically used for monitoring change detection [46]. In CUSUM, the differences between any two adjacent values (in our case, the absolute errors between any two adjacent saved model states) are measured and are compared with a threshold. CUSUM is computed as the number of the difference values higher than a threshold (called out-of-control points). In our experiment, a series of CUSUM values were generated for each model using multiple thresholds. The normalized sum of the CUSUM values, which is the smaller the better, was used for model stability evaluation.

D. *Performance Evaluation Results on Generator*

The absolute errors on all the tested radiomic features are shown in Table I. For the detailed feature-based errors, see Figure S7. On the texture features, the mean absolute error of histogram matching over all six features is 0.37. cGAN reduces it to 0.13, and GANai further reduces the absolute error significantly to 0.08 (two sample t-test $p-value \leq 0.01$). On the intensity histogram features, GANai decreases the absolute errors by 17.77% from cGAN, and 79.05% from histogram matching. The results indicate that GANai is significantly better than cGAN and patch-based histogram matching.

TABLE I: Averaged absolute errors (SD) of (1) the texture features and (2) the intensity histogram features computed using histogram matching, cGAN, and GANai. In all of them, GANai has the smallest errors (cGAN and GANai two sample t-test $p-value \leq 0.01$).

Absolute Error	**Hist. Matching**	**cGAN**	**GANai**
Contrast[1]	0.21 ±0.15	0.12 ±0.08	**0.09** ±0.06
Correlation[1]	0.18 ±0.13	0.18 ±0.12	**0.09** ±0.07
Dissimilarity[1]	0.15 ±0.11	0.09 ±0.06	**0.06** ±0.04
Energy[1]	0.47 ±0.28	0.19 ±0.14	**0.14** ±0.11
Entropy[1]	0.09 ±0.06	0.02 ±0.01	**0.01** ±0.01
Homogeneity[1]	0.28 ±0.16	0.10 ±0.06	**0.07** ±0.05
Kurtosis[2]	0.54 ±0.27	0.18 ±0.14	**0.15** ±0.11
Skewness[2]	0.51 ±0.27	0.16 ±0.12	**0.14** ±0.11

Table II shows the relative errors of GANai and cGAN on seven sets of the validation data generated using different combinations of CT acquisition parameters. A positive value indicates the error of GANai is lower than cGAN, while a negative value indicates the error of GANai is higher than cGAN. The results show that GANai outperforms cGAN on five out of seven validation subsets, on which GANai decreased the relative errors by 36.21% on average. For example, on the texture features, GANai reduces the relative error by 54.48% on the Bl64 kernel with 0.5mm slice thickness images. For the detailed feature-based errors, see Figure S8-S15.

TABLE II: Averaged relative errors on the texture features[1] and the intensity histogram features[2] by comparing cGAN and GANai. Positive values mean GANai is better, and negative values mean cGAN is better. Overall, GANai has smaller errors than cGAN.

Relative Error	Bl57 0.5mm	Bl57 1mm	Bl57 1.5mm	Bl57 3mm	Bl64 0.5mm	Bl64 1.5mm	Bl64 3mm	Overall
Contrast[1]	-0.16	0.00	0.13	0.36	0.42	-0.46	0.34	**0.25**
Correlation[1]	0.06	0.38	0.37	0.45	0.68	-0.10	0.38	**0.50**
Dissimilarity[1]	-0.05	0.28	0.32	0.48	0.64	-0.30	0.39	**0.33**
Energy[1]	-0.19	0.35	0.34	0.55	0.11	-1.61	-0.12	**0.26**
Entropy[1]	-0.05	0.30	0.30	0.48	0.67	-0.81	0.19	**0.50**
Homogeneity[1]	-0.34	0.29	0.30	0.48	0.75	-0.85	0.26	**0.30**
Kurtosis[2]	-0.05	0.18	0.26	0.45	-1.66	-1.84	-0.48	**0.17**
Skewness[2]	-0.10	0.17	0.47	0.35	-1.66	-1.84	-0.18	**0.13**

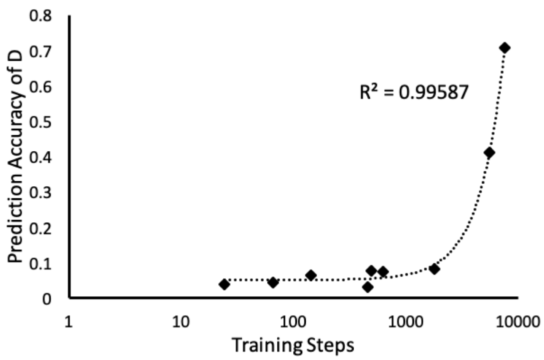

Fig. 8: Prediction accuracy of Ds gradually increases during the alternative training process of GANai.

Figure 6 shows an example of the synthesized images using cGAN or GANai generated after 100 training epochs. The GANai synthesized image is more similar to the target image, has sharper edges, and has fewer artifacts than cGAN. Figure 7 shows both cGAN and GANai model reaches their best performance after 20 epochs of training. After that, GANai can still maintain high synthesized image quality, but cGAN started to introduce artifacts.

E. Performance Evaluation Results on Discriminator

To evaluate the performance on the discriminator D, we generated a fake-pair-only dataset and used it to measure the prediction accuracy of all the Ds in the model training process. Specifically, given a fixed source image set X_{val} and the correspondent target image set Y_{val}, each having 1,750 images, we generated the synthesized image set X'_{val} using the second last generator of GANai. The accuracy of every discriminator (such as D_0 to D_3 in Figure 3) in the alternative training process of GANai was measured with all image pairs in X_{val} and X'_{val}. Accuracy is defined as the proportion of (x, x') that were correctly classified as the fake image pairs. Figure 8 shows the prediction accuracy of D at every training process. The increasing prediction accuracy shows the performance of D was steadily improving during the training of GANai.

F. Performance Evaluation Results on Training Stability

In GANai, an ensemble learning-based approach is adopted to increase the training stability. To demonstrate the effectiveness of this approach, we designed the following experiment. Three networks (cGAN, $GANai_{singleDG}$, and GANai) were trained for 100 epochs using the same training data, where $GANai_{singleDG}$ is a simplified version of GANai that trains the current component only based on the previous counter component, without using multiple Ds or Gs. The training state of every 2.5 training epochs was saved. We compared all the three models using the same validation data at every saved model state (Figure 9A). The normalized sum of the CUSUM values of cGAN, $GANai_{singleDG}$, and GANai over all the six texture features are 0.21, 0.15, and 0.13 respectively, indicating GANai is the most stable model among the three. Figure 9 shows the CUSUM on the contrast feature computed using the gray-level co-occurrence matrix.

V. Discussion

A. Training Effectiveness

The training data in GANai are separated into two subsets for the training of G and D. Our assumption is that for certain source images that are difficult to standardize, we should avoid them in the G-training phase. Instead, we use them to train D. To test the assumption, we trained a new GANai model called $GANai_{reverse}$ with the opposite training data assignment (i.e., G trained with T_{hard} and D trained with T_{easy}). Figure 10 shows that the mean absolute errors of $GANai_{reverse}$ are significantly higher than GANai on a majority of the features, indicating that training data assignment is critical for improving GAN performance.

We further tested the effectiveness of the new strategies developed for improving training effect. Two modified c-GAN models were trained, one with dedicated training data, i.e., T_{hard} for D and T_{easy} for G, called $cGAN_{SpDa}$, and the other further adopting the alternative training strategy, called $cGAN_{SpDa+AI}$. Experimental results show that 1) $cGAN_{SpDa}$ can effectively reduce the feature-based absolute errors of $cGAN$ on a majority of the texture features, and 2) $cGAN_{SpDa+AI}$ can further reduce the absolute errors on texture features (Figure 11). It indicates that the new training strategies developed in GANai are effective and can be adopted by generic GAN models to further improve their performance.

B. Effectiveness of Ensemble Learning

GANai adopts the alternatively improving strategy to train D and G so that both modules can be optimized in each iteration of training. One potential problem of such full optimization is that the model could be trapped at the local minima instead of reaching the global optimization. One such example is shown in Figure S14, where a generator has been trained for more than five epochs, but it still did not result in any significant improvement. It is reasonable to believe that the model was trapped at a local minima. To address this issue, we adopt the ensemble learning approach, i.e., GANai requires a D to discriminate multiple Gs and a G to fool multiple Ds. Also, we rollback to the previous training phase and then retrain the model, if a satisfactory loss cannot reach in a reasonable amount of time.

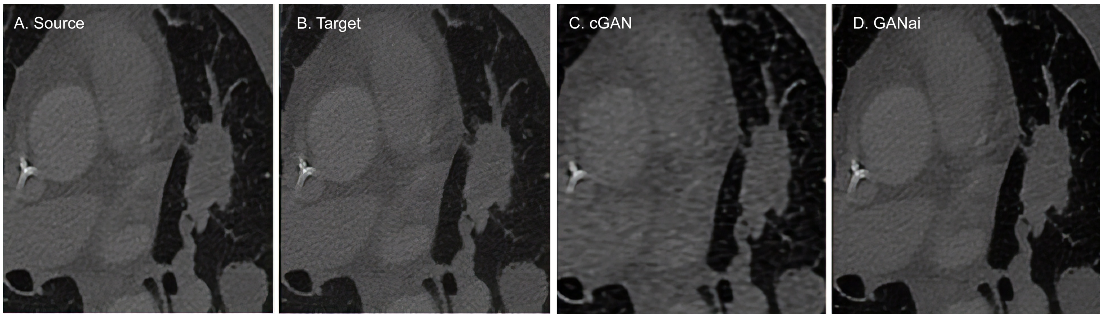

Fig. 6: Examples of the synthesized images generated by cGAN and GANai at 100th epoch. (A) source image. (B) target image. (C) cGAN synthesized image. (D) GANai synthesized image.

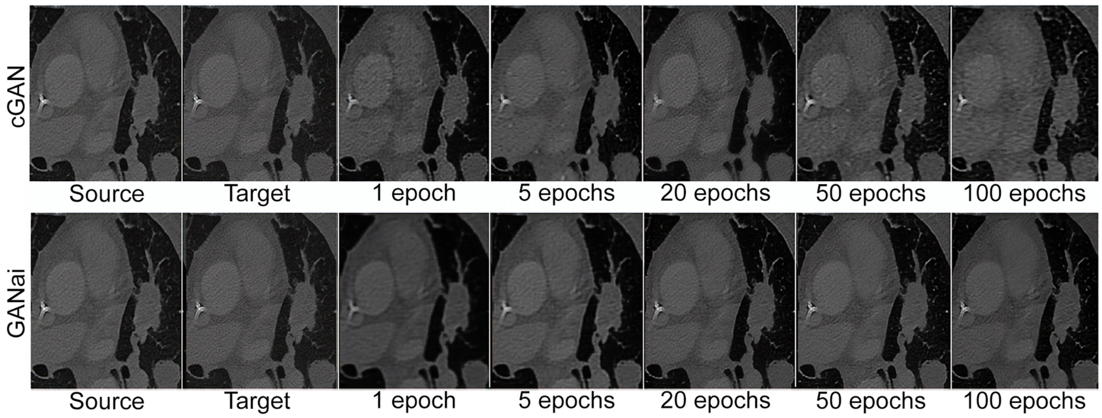

Fig. 7: Examples of the synthesized images generated by cGAN and GANai at multiple training steps. The first two columns are the source and target images. Both cGAN and GANai reached their best performance at about 20 training epochs. The synthesized images generated by cGAN have obvious artifacts and have less sharp edges than that of GANai. Furthermore, GANai maintained a high synthesized image quality in the continuous training after the first 20 epochs, whereas cGAN started to introduce additional artifacts into the synthesized images.

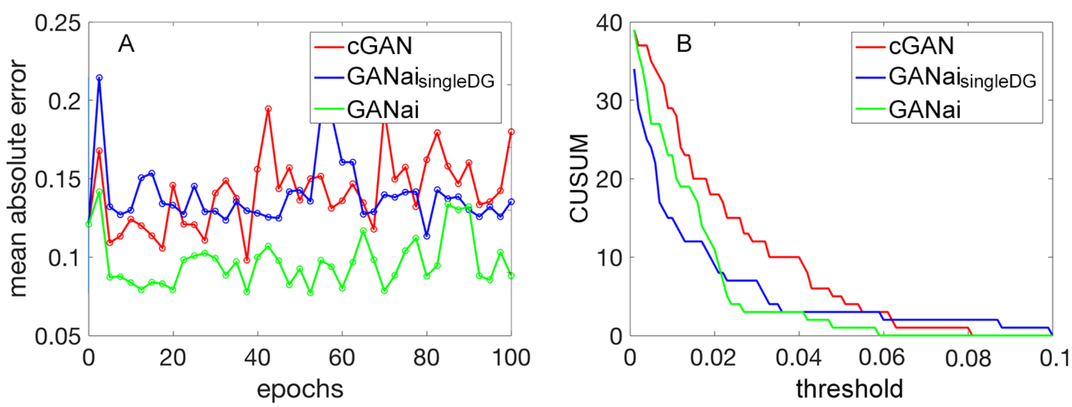

Fig. 9: Performance evaluation on training stability. (A) the mean absolute errors of cGAN, $GANai_{singleDG}$, and GANai on the contrast feature computed using the gray-level co-occurrence matrix. (B) the CUSUM values of cGAN, $GANai_{singleDG}$, and GANai, where the x-axis is the threshold of CUSUM, and the y-axis is the CUSUM value. In general, GANai is the most stable model among the three.

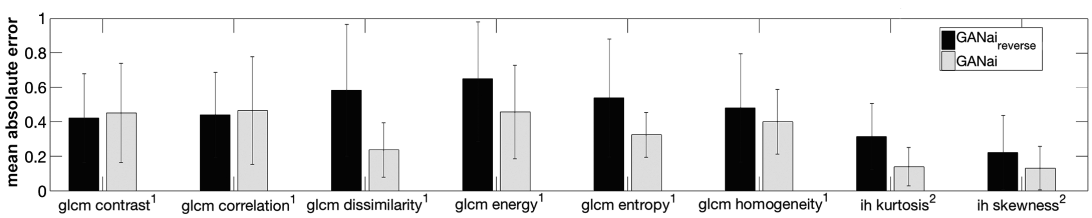

Fig. 10: Averaged feature errors for the data effectiveness test. [1] Gray-level co-occurrence matrix, [2] Intensity Histogram. It shows that the mean absolute errors of $GANai_{reverse}$ are significantly higher than GANai on a majority of the features, indicating that training data assignment is critical for improving GAN performance.

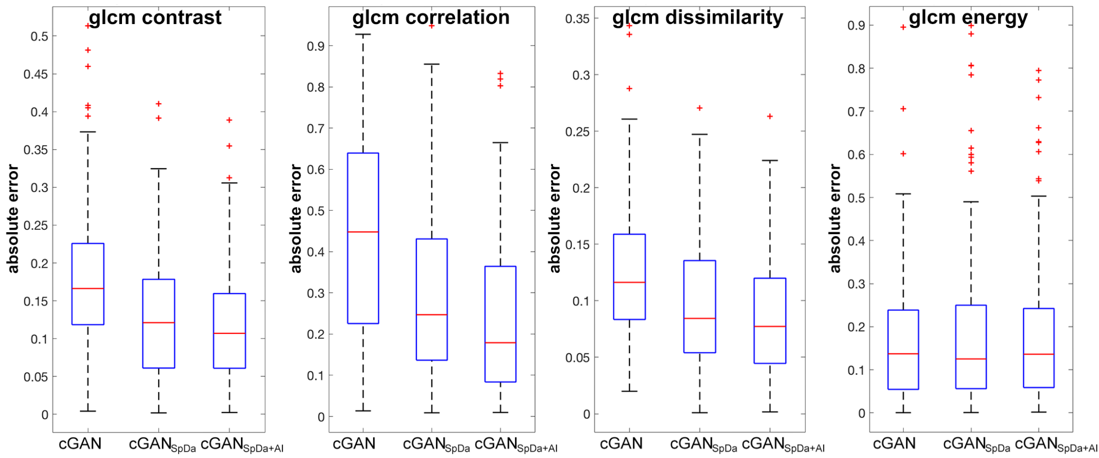

Fig. 11: Gray-level co-occurrence matrix feature errors of different cGAN versions. $cGAN_{SpDa}$ was trained with dedicated training data. $cGAN_{SpDa+AI}$ was further adopting the alternative training strategy.

C. Validation Loss

The validation loss of G in Figure 5B is constantly lower than the training loss, which is uncommon to machine learning tasks. This is reasonable because the loss of G is $-logD(x, x')$ computed using the prediction result on all the fake image pairs. As shown in Figure 4, the value of $D(x, x')$ on the validation dataset is higher than that on the training dataset. After taking the minus log, the validation loss is smaller than the training loss. However, as stated in Gulrajani et al [47], the loss of GANs may not associate with model performance. Thus, the fact that the validation loss of G is smaller than the training loss does not necessarily indicate whether the synthesized images on the validation dataset is better than that on the training dataset. It is also why GANai uses the prediction of D, rather than using the loss of G, to control the model training phase switch.

D. Limitations

While GANai, in general, performs better than traditional GAN models and histogram matching on texture features, its performance could be suboptimal on shape-based features. Shape-based features, such as volume, are usually determined by the physical setup of CT machines. For instance, a 1.5 mm nodule can be totally omitted in a 3 mm slice thickness scan due to partial volume [48].

VI. Conclusion

As a popular diagnostic image modality, CT is routinely used for assessing anatomical tissue characteristics. However, CT imaging customization poses a fundamental challenge in radiomics, since non-standardized imaging protocols are commonplace. Image synthesis algorithms have been developed to integrate and standardize CT images. Among them, GAN models learn the data distribution of training data and generate synthesized images under the same distribution of the training images. However, GANs are not directly applicable to the CT image mitigation task due to the lack-of-detail problem.

We developed a novel GAN model called GANai to mitigate the differences in radiomic features of CT images. Given source images, GANai composes synthesized images by specifying a high-level goal that the image features of the synthesized images should be similar to those of the target images. GANai introduces the alternative training strategy to GAN. In each training phase, the model aims to optimize either G or D while freezing the other component. A training phase will stop if the current component is well trained or the training step exceeds an upper bound. After that, GANai switches to train the counter component. Note that just because of the adoption of the alternative training strategy, new technical improvements become applicable. For example, the inputs of the ensemble learning (multiple states of Ds and Gs) are the end products of every alternative training phase, and a new loss function and dedicated training data can be specified in different training phases. GANai was compared with the start-of-the-art cGAN model [22] and the patch-based histogram matching method [16]. The experimental results show that GANai is significantly better than cGAN and patch-based histogram matching on the texture and intensity histogram based radiomic features.

In conclusion, GANai is a new GAN model for CT image standardization. Its alternative training strategies are effective, easy to implement, and can be adopted by the other GAN models to further improve their performance. With GANai, CT images from multiple medical centers can be seamlessly integrated and standardized, and large-scale radiomics studies can be conducted to extract comprehensive radiomic features and to identify key tumor characteristics that drive disease transformation, progression, and drug resistance.

References

[1] J. L. Prince and J. M. Links, *Medical imaging signals and systems*. Pearson Prentice Hall Upper Saddle River, 2006.

[2] J. Beutel, H. L. Kundel, and R. L. Van Metter, *Handbook of medical imaging: Physics and psychophysics*. Spie Press, 2000, vol. 1.

[3] A. Webb and G. C. Kagadis, "Introduction to biomedical imaging," *Medical Physics*, vol. 30, no. 8, pp. 2267–2267, 2003.

[4] M. Mahesh, "Fundamentals of medical imaging," *Medical Physics*, vol. 38, no. 3, pp. 1735–1735, 2011.

[5] J. T. Bushberg and J. M. Boone, *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.

[6] A. Midya, J. Chakraborty, M. Gönen, R. K. Do, and A. L. Simpson, "Influence of ct acquisition and reconstruction parameters on radiomic feature reproducibility," *Journal of Medical Imaging*, vol. 5, no. 1, p. 011020, 2018.

[7] S. P. Raman, M. Mahesh, R. V. Blasko, and E. K. Fishman, "Ct scan parameters and radiation dose: practical advice for radiologists," *Journal of the American College of Radiology*, vol. 10, no. 11, pp. 840–846, 2013.

[8] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.

[9] A. J. Buckler, L. Bresolin, N. R. Dunnick, D. C. Sullivan, and Group, "A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging," *Radiology*, vol. 258, no. 3, pp. 906–914, 2011.

[10] G. Liang, J. Zhang, M. Brooks, J. Howard, and J. Chen, "radiomic features of lung cancer and their dependency on ct image acquisition parameters: su-k-201-12," *Medical Physics*, vol. 44, no. 6, p. 3024, 2017.

[11] R. Berenguer, M. d. R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. M. Legorburo, and S. Sabater, "Radiomics of ct features may be nonreproducible and redundant: Influence of ct acquisition parameters," *Radiology*, p. 172361, 2018.

[12] L. A. Hunter, S. Krafft, F. Stingo, H. Choi, M. K. Martel, S. F. Kry, and L. E. Court, "High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images," *Medical physics*, vol. 40, no. 12, 2013.

[13] M. F. Cohen and J. R. Wallace, *Radiosity and realistic image synthesis*. Elsevier, 2012.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 1060–1069.

[15] A. Ben-Cohen, E. Klang, S. P. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. M. Amitai, and H. Greenspan, "Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection," *arXiv preprint arXiv:1802.07846*, 2018.

[16] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, NJ: Prentice Hall, 2012.

[17] A. R. Weeks, L. J. Sartor, and H. R. Myler, "Histogram specification of 24-bit color images in the color difference (cy) color space," *Journal of electronic imaging*, vol. 8, no. 3, pp. 290–301, 1999.

[18] A. Rosenfeld, *Digital picture processing*. Academic press, 1976.

[19] A. K. Jain, *Fundamentals of digital image processing*. Englewood Cliffs, NJ: Prentice Hall,, 1989.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784v1*, 2014.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] B. S. Rosenstein, C. M. West, S. M. Bentzen, J. Alsner, C. N. Andreassen, D. Azria, G. C. Barnett, M. Baumann, N. Burnet, J. Chang-Claude *et al.*, "Radiogenomics: radiobiology enters the era of big data and team science," *International Journal of Radiation Oncology* Biology* Physics*, vol. 89, no. 4, pp. 709–713, 2014.

[24] Q. Li, J. Kim, Y. Balagurunathan, Y. Liu, K. Latifi, O. Stringfield, A. Garcia, E. G. Moros, T. J. Dilling, M. B. Schabath *et al.*, "Imaging features from pre-treatment ct scans are associated with clinical outcomes in non-small-cell lung cancer patients treated with stereotactic body radiotherapy," *Medical physics*, vol. 44(8), pp. 4341–4349, 2017.

[25] H. J. Aerts, "The potential of radiomic-based phenotyping in precision medicine: a review," *JAMA oncology*, vol. 2, no. 12, pp. 1636–1642, 2016.

[26] D. V. Fried, S. L. Tucker, S. Zhou, Z. Liao, O. Mawlawi, G. Ibbott *et al.*, "Prognostic value and reproducibility of pretreatment ct texture features in stage iii non-small cell lung cancer," *International Journal of Radiation Oncology* Biology* Physics*, vol. 90, no. 4, pp. 834–842, 2014.

[27] K. Thayalan and R. Ravichandran, *The physics of radiology and imagings*. Jaypee Brothers Medical Publishers, 2014.

[28] A. N. Primak, C. H. McCollough, M. R. Bruesewitz, J. Zhang, and J. G. Fletcher, "Relationship between noise, dose, and pitch in cardiac multi–detector row ct," *Radiographics*, vol. 26, no. 6, pp. 1785–1794, 2006.

[29] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, no. 542, pp. 115–118, February 2017.

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[31] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[32] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, International Convention Centre, Sydney, Australia, 2017, pp. 1857–1865.

[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[34] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," *arXiv preprint arXiv:1803.07422*, 2018.

[35] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[36] Y. Wang, L. Zhang, and J. van de Weijer, "Ensembles of generative adversarial networks," *arXiv preprint arXiv:1612.00991*, 2016.

[37] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," *arXiv preprint arXiv:1611.01673*, 2016.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[39] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv:1610.01945*, 2017.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[41] A. Strehl and J. Ghosh, "Cluster ensembles – a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[43] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1, 2015, pp. 1486–1494.

[44] I. G. Tim Salimans, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2234–2242.

[45] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[46] O. A. Grigg, V. Farewell, and D. Spiegelhalter, "Use of risk-adjusted cusum and rsprtcharts for monitoring in medical contexts," *Statistical methods in medical research*, vol. 12, no. 2, pp. 147–170, 2003.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv:1704.00028*, 2017.

[48] A. A. Divani, S. Majidi, X. Luo, F. G. Souslian, J. Zhang, A. Abosch, and R. P. Tummala, "The abcs of accurate volumetric measurement of cerebral hematoma," *Stroke*, vol. 42, no. 6, pp. 1569–1574, 2011.