Journal of Perceptual Imaging® 5: 000502-1–000502-15, 2022.

Controllable Medical Image Generation via GAN

Zhihang Ren[1,2], Stella X. Yu[1,2], David Whitney[1,2,3,4]

[1] *Vision Science Graduate Group, University of California, Berkeley, CA 94720, United States of America;* [2] *International Computer Science Institute, Berkeley, CA 94720, United States of America;* [3] *Department of Psychology, University of California, Berkeley, CA 94720, United States of America;* [4] *Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, United States of America*

E-mail: peter.zhren@berkeley.edu

Abstract. *Medical image data is critically important for a range of disciplines, including medical image perception research, clinician training programs, and computer vision algorithms, among many other applications. Authentic medical image data, unfortunately, is relatively scarce for many of these uses. Because of this, researchers often collect their own data in nearby hospitals, which limits the generalizabilty of the data and findings. Moreover, even when larger datasets become available, they are of limited use because of the necessary data processing procedures such as de-identification, labeling, and categorizing, which requires significant time and effort. Thus, in some applications, including behavioral experiments on medical image perception, researchers have used naive artificial medical images (e.g., shapes or textures that are not realistic). These artificial medical images are easy to generate and manipulate, but the lack of authenticity inevitably raises questions about the applicability of the research to clinical practice. Recently, with the great progress in Generative Adversarial Networks (GAN), authentic images can be generated with high quality. In this paper, we propose to use GAN to generate authentic medical images for medical imaging studies. We also adopt a controllable method to manipulate the generated image attributes such that these images can satisfy any arbitrary experimenter goals, tasks, or stimulus settings. We have tested the proposed method on various medical image modalities, including mammogram, MRI, CT, and skin cancer images. The generated authentic medical images verify the success of the proposed method. The model and generated images could be employed in any medical image perception research.* © *2022 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.Percept.Imaging.2022.5.000502]

1. INTRODUCTION

Medical imaging has transformed modern medicine, allowing clinicians to noninvasively examine and diagnose patients with remarkable ease and speed. In recent years, there have been dramatic advances in the field of medical imaging technologies, ranging from MRI, CT, PET, photography, ultrasound, among many other techniques. These improvements are astounding, but it is worth noting that ultimately the data provided by these techniques requires critical human involvement in detection, selection, interpretation, and diagnosis. The imaging techniques themselves are not the only bottleneck for obtaining accurate diagnoses.

Fortunately, along with the technological developments, there have also been concomitant advances in the application and use of these technologies. For instance, there is a recent surge in computer vision and medical image perception research, that require artificial (algorithmic) and human users respectively. In both machines and humans, there is a great deal of potential to improve the use of medical imaging in clinical practice. In addition to the more ambitious goals of automated diagnoses, filtering, or cuing clinicians [34, 51, 68, 86], there are distinct and more pressing goals of improving clinicians' medical image perception and decision-making [74, 78, 81] in the realms of training, error detection, diagnostic support, among others [79].

To improve both machine and human medical image perception, it is necessary to have sufficient source data. Unfortunately, labeled and de-identified public medical imaging data is scarce. Sometimes researchers resort to collecting their own data from nearby hospitals, usually from local areas that cannot represent the broader population. Second, even if larger datasets are collected, the necessary data processing procedures such as data de-identification, labeling, and categorizing requires significant time and effort. For instance, in certain medical imaging tasks, such as lesion segmentation, in order to prepare the training data, it requires experts to perform meticulous annotations that are tedious and labor intensive [80]. Moreover, collected medical images are specific to each individual patient and it can be difficult to find specific images or image properties that satisfy certain desired experimental configurations [52]. Of course, due to intricate tissue structures, manipulating attributes of those collected medical images using traditional image processing methods is difficult or impossible, at least in a realistic manner.

The data scarcity problem has presented a major challenge to research on medical image perception. At a broad level, medical image perception research studies the visual and cognitive processes that clinicians rely on to make

Received June 15, 2021; accepted for publication Jan. 10, 2022; published online March 18, 2022. Associate Editor: Damon Chandler.

2575-8144/2022/5/000502/15/$00.00

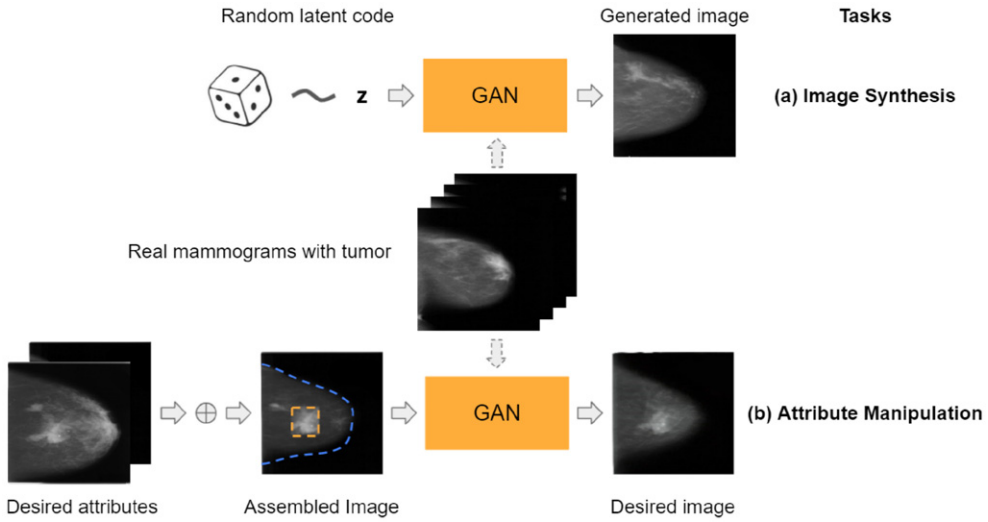

Figure 1. Pipeline. Controllable medical image generation using the proposed GAN model. (a) Medical image generation: novel and authentic medical images can be generated from random latent codes z. (b) Attribute manipulation: desired attributes can be assembled together to satisfy certain experimental settings. Here, we use mammogram as an example medical modality. Real mammograms with tumor were utilized to train the proposed model. Our proposed model can be easily adapted to other medical modalities, such as MRI, CT, and skin cancer images.

decisions. As in other domains of human factors, the goal of understanding those mechanisms is to improve (i.e., guide, cue, facilitate, speed, etc) clinician performance. Recently, in many psychophysical experiments, artificial medical stimuli have been employed [46, 52]. The artificial medical stimuli are often composed of simple shapes or textures with some form of noise background [46, 52, 77]. Related approaches involve using real medical images but superimposing clearly artificial "targets" [46, 52]. An advantage of these approaches is that they are relatively easy to generate and control in a precise manner, which is important for studying the cognitive and perceptual systems of clinicians [46, 52]. For example, the image attributes and "targets" are easy to manipulate such that researchers can perform shape morphing and background replacement. This level of stimulus control is necessary in perception research to study things like visual search for lesions, visual recognition of lesions, inattentional blindness, cognitive load and interference, etc. However, those artificial medical images are obviously inauthentic, completely unlike what clinicians routinely examine. Thus, the results of these experiments fall invariably within a shadow of a doubt about clinical applicability.

Therefore, generating authentic and easily controllable medical images is critical for the entire field of medical image perception research. Alleviating constraints is only recently realistic, with the impressive development of deep learning in computer vision. For example, Generative Adversarial Network (GAN) is one of the promising models that have achieved great success on image generation tasks. GAN can generate high-quality authentic images with various categories [41, 42, 60], such as faces, cars, landscapes, and so on. Additionally, various methods can be applied to manipulate the attributes of Generative Adversarial Networks' outputs [13, 54, 60].

In this paper, we utilize Generative Adversarial Network (GAN) to generate authentic medical images (Figure 1(a)). We also adopt a controllable approach to manipulate specific attributes of the generated images (Fig. 1(b)). The proposed method is tested on various medical image modalities such as mammogram, MRI, CT, and skin cancer images. For example, via controllable generation, we can create authentic mammograms with desired tumor and breast shapes. We also recruited both expert clinicians and untrained participants to discriminate the authenticity of each image (real versus GAN generated) in an objective psychophysical experiment. Finally, we investigate the perceptual loss which is utilized in the controllable generation. Various experiments verify the success of the proposed controllable medical image generation model.

Contributions: We propose a framework for controllable medical image generation with the following contributions.

- We propose to utilize Generative Adversarial Network (GAN) to generate medical images and verify the results on various medical image modalities such as mammogram, MRI, CT, and skin cancer images.
- We adopt a controllable approach to manipulate the attributes of the generated images in order to meet certain experimental configurations.
- We compare traditional similarity measurements with the perceptual metric in medical imaging.

Although a shorter conference version of this paper appeared in [66], it was limited in scope and did not extend the model to multiple medical image modalities. This paper extends the model to MRI, CT, and skin cancer images. Moreover, this paper compares traditional similarity measurements with the perceptual metric in medical imaging.

2. RELATED WORK

2.1 *Convolutional Neural Networks*

The idea of Convolutional Neural Networks (CNN) stem from the discovery of the edge detector in cat's striate cortex [38]. Based on this finding, Fukushima [23] invented the first simple hierarchical, multilayered artificial neural network. After decades of development, LeCun et al. [49] leveraged CNN for hand-written ZIP Code numbers recognition and trained the network end-to-end via gradient descent. This fully automatic image recognition model can be applied to many image categories and types. The great success is mainly attributed to the convolution operation, which can reveal the latent semantic information of an image, and the shared hierarchical kernels, which make the convolution shift-invariant. During training, the loss is computed based on specific metrics for certain tasks, updating the model parameters while it back propagates through the whole network.

However, the computation is heavy, which limits the model's capacity and ability for high-resolution images. With the deployment of Graphical Processing Unit (GPU), CNNs [33, 47, 71–73] have shown promise in computer vision tasks, such as image classification [33], object detection [27, 28, 64, 65], and object segmentation [32]. Recently, many medical imaging tasks have been utilizing CNNs [22, 44, 70]. Compared to traditional image processing methods, CNNs have much better performance with much faster inference speed.

2.2 *Generative Adversarial Networks*

Generative Adversarial Networks are special Convolutional Neural Networks, which consist of two networks, the generator (G) and the discriminator (D). These two networks are trained iteratively in an adversarial way where the generator (G) generates fake but authentic images to fool the discriminator and the discriminator (D) discriminates the real and fake images [29]. Using this promising computational model, high-quality images with various categories can be generated, such as faces, cars, and landscapes [41, 42, 60]. However, the initial GAN model [29] cannot generate sharp and recognizable images, and the training process is unstable. Later work improved the performance of GAN in different ways. Some papers focus on model architectures [13, 54, 58]. Others focus on improving the loss metrics and training strategies [2, 9, 30]. With these efforts, GAN training stability has improved, and GAN can generate low-resolution images with sufficient quality.

Of late, numerous approaches for high-resolution image generation are also available. PGGAN [41] aims to train the standard GAN from coarse to fine scale. The parameters for low-resolution block are trained first. Then higher-resolution blocks are added on gradually with the corresponding parameters updated accordingly. Based on the same training strategy, StyleGAN [42, 43] proposed to first map the original latent space $\mathcal{Z}$ into the $\mathcal{W}$ space through a non-linear mapping network. Then it is merged into the synthesis network via adaptive instance normalization (AdaIN) at each convolutional block [17, 36]. This improves StyleGAN representations of scenes and details and allows it to produce authentic high-resolution images. In this paper, we adopt StyleGAN as our backbone for medical images generation. Moreover, a controllable approach is also utilized to manipulate the attributes of the generated images.

In medical image applications, [22] DCGAN [61] and ACGAN [58] were utilized to generate CT liver lesion patches and boosted the liver lesion classification performance. Han et al. [31] deployed WGAN [30] to generate MR images for data augmentation and physician training. Nie et al. [57] used GAN to predict CT images from MR images. Cao et al. [10] proposed an Auto-GAN to synthesize missing modality for medical images. Moreover, GAN has been widely used for skin cancer image generation and purification [5–7, 26]. Our approach is different from aforementioned methods. In addition to purely generating new samples as GANs traditionally do, our method can also edit specific images via the encoder of our model.

2.3 *Perceptual Loss*

CNN features have already been utilized for calculating similarity for years. Ref. [1] proposed to use pre-trained AlexNet features for image quality measurement. Perceptual loss, which is also based on CNN features, was first proposed in Ref. [40] for style transfer [25] and super resolution tasks. Both are ill-posed problems. For style transfer, there is no absolute ground truth image for reference. For image super resolution, one low-resolution image can have many corresponding high-resolution images which can be down-sampled to the same low-resolution image. Thus, per-pixel metric is no longer suitable since semantic similarity matters. Recently, traditional similarity metrics, such as Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), are found to be inconsistent with human perception, and a perceptual metric has been utilized to measure the semantic similarity in many papers [37, 50, 87, 89]. In this paper, we use perceptual loss to regularize the encoder training and guide the latent code optimization in the encoding procedure.

3. METHOD

Here, we adapt the Generative Adversarial Network for medical image generation. In order to manipulate the image attributes, an encoder is added to encode certain image attributes into the latent code z which is the input of the GAN generator.

Our proposed model is composed of two parts. The first part is the GAN which involves the generator (G) and the discriminator (D). The generator (G) will generate authentic (fake) images from the latent codes z, and try to fool the discriminator (D) during training. The discriminator (D) will discriminate whether the image is real (i.e. sampled from real images) or fake (i.e. generated from the generator), and try to beat the generator by distinguishing the fake images from the real ones. The second part of the model is the encoder (E), which can encode image attributes into the

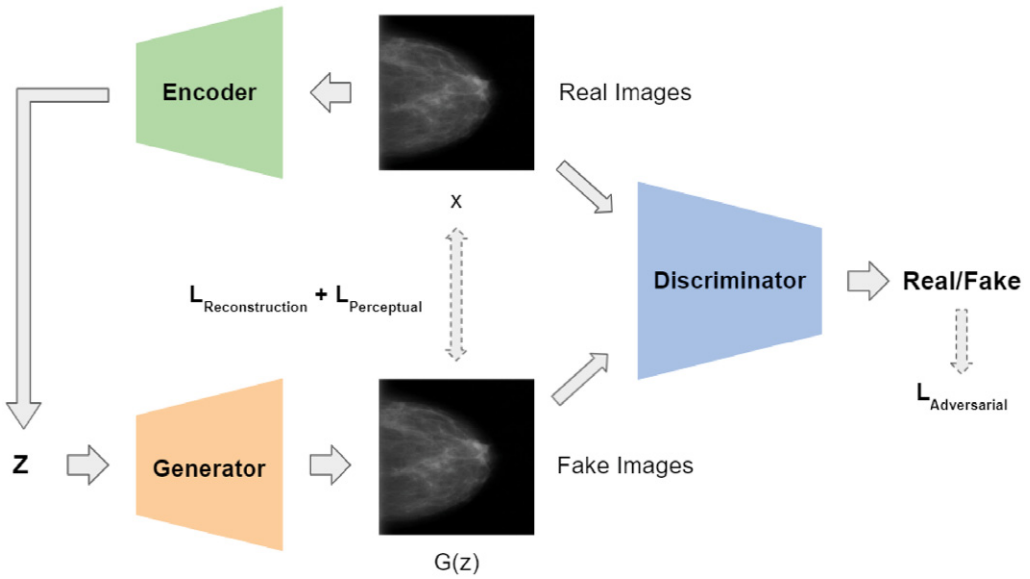

Figure 2. Architecture of proposed method. The architecture contains three sub-networks, the encoder (E), the generator (G), and the discriminator (D). The training has two phases. In the first phase, the generator and discriminator will be trained first without the encoder (E) via adversarial loss $L_{\text{adversarial}}$. In the second phase, the generator (G) will be fixed. The encoder (E) and discriminator (D) will be trained adversarially via the reconstruction loss $L_{\text{reconstruction}}$, the perceptual loss $L_{\text{perceptual}}$, and the adversarial loss $L_{\text{adversarial}}$. The dashed arrows indicate how to compute the corresponding loss metrics.

latent code z. This latent code can then be utilized to generate the image through the generator. Therefore, it can allow us to manipulate the generated image by manipulating the latent code through the encoder. The architecture is shown in Figure 2.

While training, the GAN part is first trained progressively [42] via adversarial loss $L_{\text{Adversarial}}$. The training process can be formulated as

$$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim q(z)}[\log(1 - D(G(z)))], \tag{1}$$

where $p_{\text{data}}(x)$ and $q(z)$ indicate the real data distribution and the latent space distribution respectively, x is the sampled real image, z is the sampled latent code.

Then, we train the encoder part. After training the GAN part, the generator (G) is fixed. While training the encoder network, traditional methods [3] regularize the encoder on the latent space, encouraging the encoder to encode the same latent codes for the corresponding generated images regardless of the reconstructed images. This method can degrade the reconstruction quality. Instead, we adopt the idea from In-domain GAN inversion [88], where the regularization of the encoder is on the image space. In particular, the encoded vector is passed into the generator (G) again and the regularization is on the reconstructed image. The L2 reconstruction loss $L_{\text{Reconstruction}}$ and the perceptual loss [40] $L_{\text{Perceptual}}$ are utilized for the regularization. Additionally, adversarial loss $L_{\text{Adversarial}}$ is also utilized to guarantee that the reconstructed image looks authentic. The whole process can be summarized as follows

$$\min_E \|x - G(E(x))\|_2 + \lambda_1 \|F(x) - F(G(E(x)))\|_2 - \lambda_2 E_{x\sim p_{\text{data}}(x)}[\log D(G(E(x)))], \tag{2}$$

$$\min_D E_{x\sim p_{\text{data}}(x)}[\log D(G(E(x)))] - E_{x\sim p_{\text{data}}(x)}[\log D(x)] + \frac{\gamma}{2} E_{x\sim p_{\text{data}}(x)}[\|\nabla_x D(x)\|_2^2], \tag{3}$$

where $p_{\text{data}}(x)$ indicates the real data distribution, x is the real image, E represents the encoder, F represents the VGG feature extraction [71], and λ_1, λ_2 and γ are weights for the perceptual loss, the adversarial loss, and the gradient penalty [30].

Since the inverse mapping via the encoder (E) will not always be perfect, in order to get the optimal inverse latent code, we apply another optimization on the latent code. This optimization will update the latent code based on the reconstruction loss and the perceptual loss within the neighborhood of the original encoded vector (the encoder regularization). The optimization process can be described as below

$$z^{\text{inv}} = \min_z \|x - G(E(x))\|_2 + \lambda_3 \|F(x) - F(G(z))\|_2 + \lambda_4 \|z - E(G(z))\|_2, \tag{4}$$

where z^{inv} is the optimized inverse code, λ_3 and λ_4 are weights for the perceptual loss, and the code reconstruction loss (i.e., the encoder regularization). This optimization metric can be computed using the whole image region (for image reconstruction) or the region of interest (for image manipulation).

3.1 *Medical Image Synthesis*

In general, informative images lie on a manifold. Through the GAN training, the generator (G) learns a transformation from the latent space to the image space, imitating the real

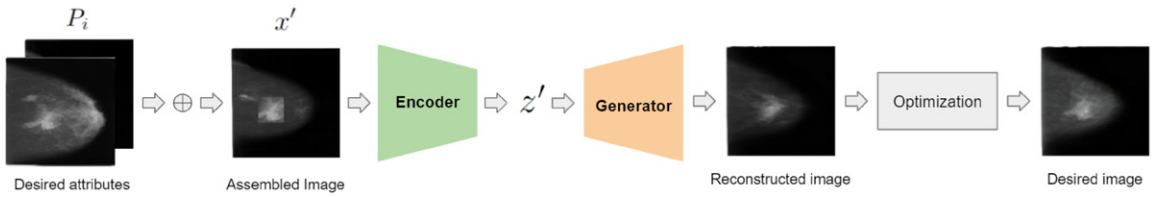

Figure 3. Attribute manipulation pipeline. Firstly, the desired image attributes are combined by merging image patches that contain those attributes. Then, the corresponding latent code is produced by the encoder. The generator reconstructs the image with desired attributes. Finally, the desired image can be obtained after the final optimization.

image manifold of the training dataset. Thus, we can utilize this learned transformation to generate images authentic to the real images. First, the latent code z will be sampled from the latent space. Then, the output image $x = G(z)$ is produced by the generator.

Using the learned transformation, we can also generate similar medical images. As a manifold, the nearby images on the manifold are similar to each other. Therefore, we can sample a series of latent codes z_i on a closed path C, then passing these latent codes into the generator (G), we can obtain a series of gradually and continuously morphing images x_i:

$$x_i = G(z_i), z_i \sim C. \tag{5}$$

3.2 *Attribute Manipulation*

While training the encoder (E), without the discriminator (D), the encoder and the generator form an autoencoder [53, 63]. The training encourages the encoder to embed useful image attributes into the latent code. Since the generator is pretrained under the GAN, the generator has learned how to reconstruct the embedded image attributes with proper tissue context.

In order to manipulate the image attributes, we first need to combine the desired image attributes into one assembled image x'. The combination can be achieved by merging image patches P_i which contain the desired image attributes:

$$x' = \bigcup_{i=1}^{n} P_i. \tag{6}$$

Then this assembled image x' will be encoded by the encoder, $z' = E(x')$, obtaining the corresponding image attributes latent code z'. The generator will finally reconstruct those image attributes with proper tissue texture, $x_{\text{reconstruct}} = G(z')$.

Since the image with all desired attributes may not exist on the image manifold, the reconstructed image may not have the exact desired attributes as we designed. The final optimization (shown in Eq. (4)) can be conducted on the region where the attributes need to be accurate. The pipeline for attribute manipulation is shown in Figure 3.

4. EXPERIMENTS AND RESULTS

4.1 *Implementation Details*

For the Generative Adversarial Network (GAN), we adopt StyleGAN [42]. The training is progressive. Starting from 8×8, the latter resolution blocks are added progressively after the previous blocks finish training. The output image resolution is 256×256. While training the encoder, the generator is fixed. Only the encoder and discriminator parameters are updated. For the perceptual loss, VGG [71] *conv*4_3 feature layer is utilized. As for the hyperparameters, $\lambda_1 = 0.00005$, $\lambda_2 = 0.1$, $\lambda_3 = 0.00005$, $\lambda_4 = 2$, and $\gamma = 10$. We use the Adam optimizer [45] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is set to 0.0001. Pytorch is utilized for coding.

For mammogram images, we use DDSM [8] dataset which contains 2, 620 normal, benign, and malignant cases. Only the benign and malignant cases are utilized for training. For MRI images, we utilize fastMRI [84] multi-coil dataset which contains 7135 images. For CT images, DeepLesion [83] dataset is used. We utilize the abdomen image dataset which contains 14601 images. For skin cancer images, we use images from ISIC Archive (https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main) which contains 69445 images in total.

4.2 *GAN Generated Results*

For different medical image modalities, we train the whole network separately using corresponding datasets. After the GAN part has been trained, we randomly sample latent codes z and pass them to the generator. The generated results for Mammogram, MRI, CT, and Skin Cancer are shown in Figure 4. Compared to the real samples on the left, the generated samples on the right appear very similar, and this is seen across different medical image modalities. It is clear that the generator has learned the semantic statistics of the training dataset for different medical image modalities. The generator can generate authentic tissue texture, tissue distribution, tissue shapes, and color distributions. Moreover, the generator can not only reconstruct the original medical images, but also it can produce novel and authentic medical images which do not actually exist in the real world.

Since the GAN training learns the manifold of the training dataset, we can also generate gradually and continuously morphing medical images for certain experiments. First, the latent codes need to be sampled from a closed path in the latent space. To do so, we randomly pick three anchor points in the latent space and calculate the interpolations between each pair of them. Then, passing those codes to the generator, we can obtain the gradually and continuously morphing medical images. The result is shown in Figure 5. Due to the

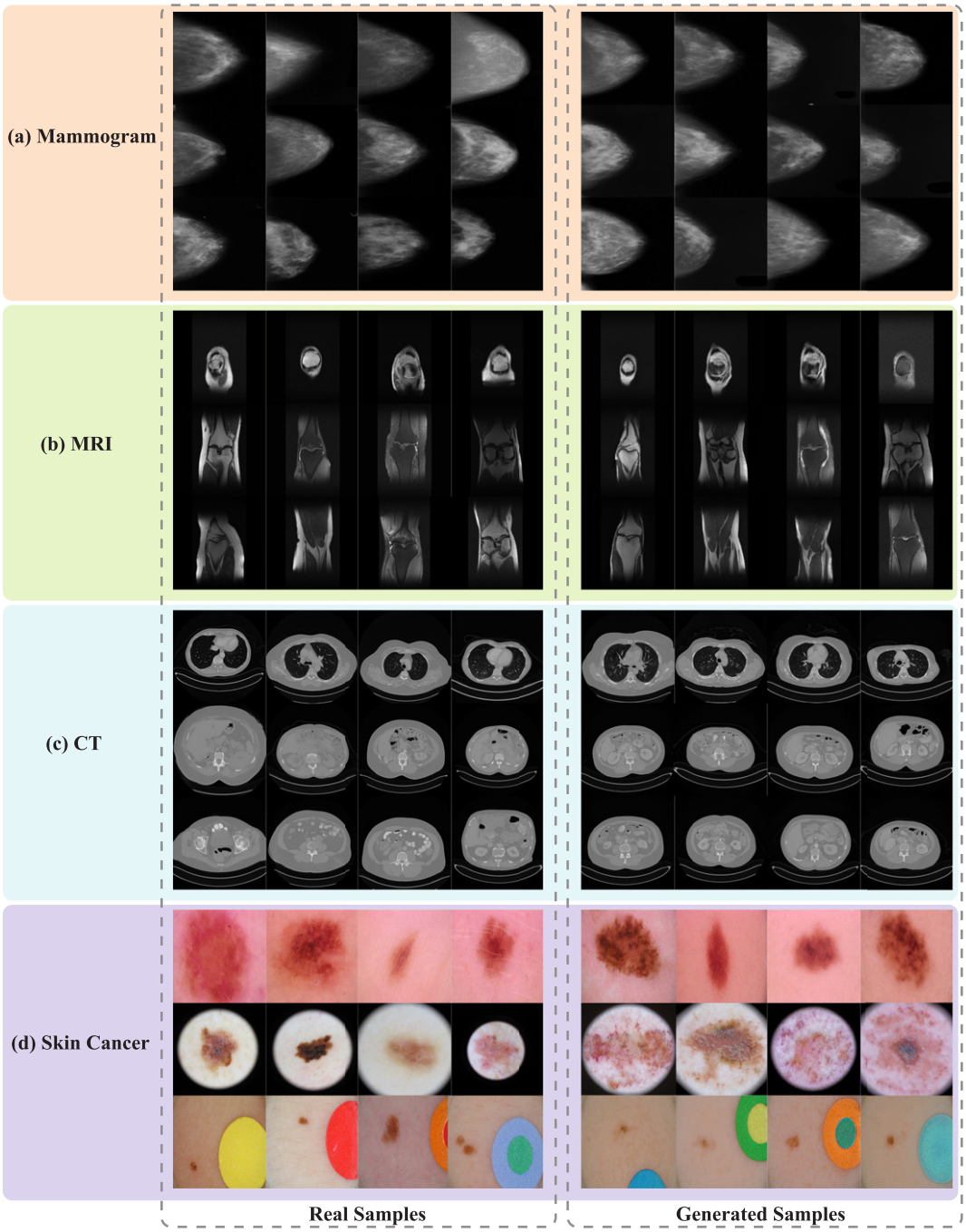

Figure 4. GAN generated results. The generated results for different medical image modalities. Comparing the real samples to the generated samples, it is clear that the generator has learned how to imitate tissue texture, tissue distribution, tissue shapes, and color distribution. The generator can produce authentic images (see below for psychophysical results confirming this).

space limit, we only show three interpolations between each pair; arbitrarily fine grained interpolations can be created between any number of pairs.

4.3 *Attribute Manipulation*

Our proposed model can generate desired medical images by manipulating the image attributes. For illustration, we show how we generate mammograms with the desired lesion patch and desired breast shapes. The results are shown in Figure 6.

First, we combine the desired image attributes, i.e. the lesion patch (Fig. 6A) and shape templates (Fig. 6B), by merging the lesion patch and shape templates directly. Then we encode these intermediate combined images (Fig. 6C) using the encoder and pass the codes to the generator. The reconstructed images from the generator are shown in Fig. 6(E) (without optimization). It is clear that the shapes are already the same as the shape templates and the overall texture is authentic. But the desired lesion texture is not

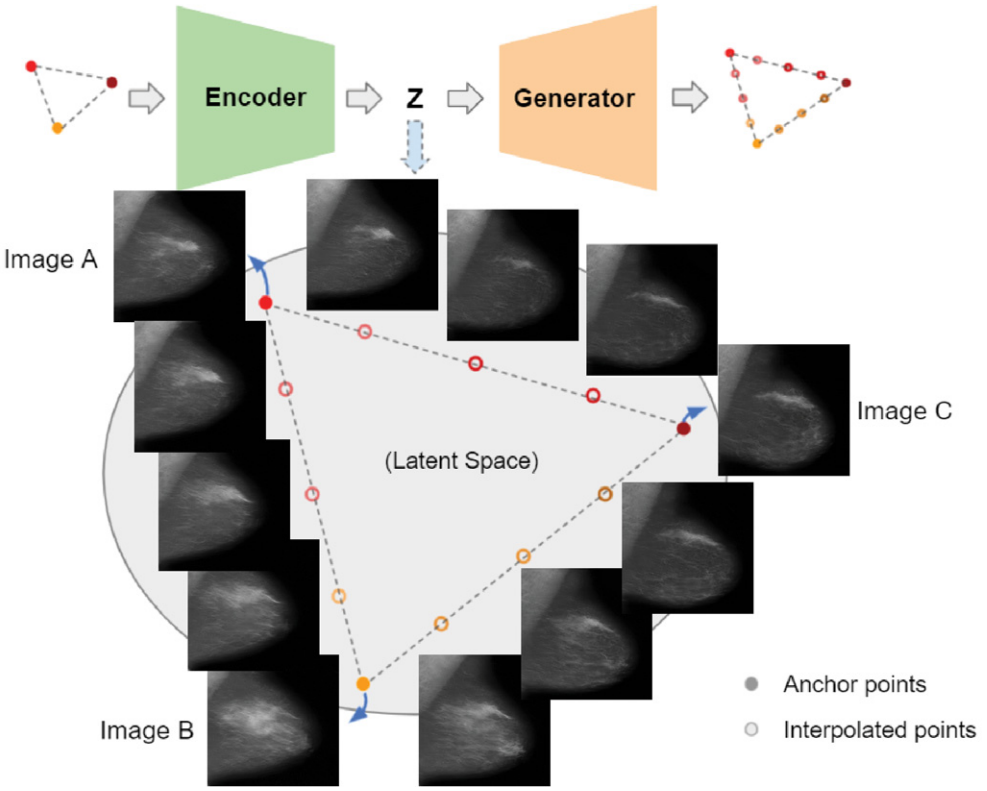

Figure 5. Interpolation results. Here, we show a mammogram loop gradually changing among three anchor images. The mammograms between two of the anchor images are generated by passing the interpolated codes of those two anchor images to the trained generator. Any number of interpolated images between any pair of anchors can be created.

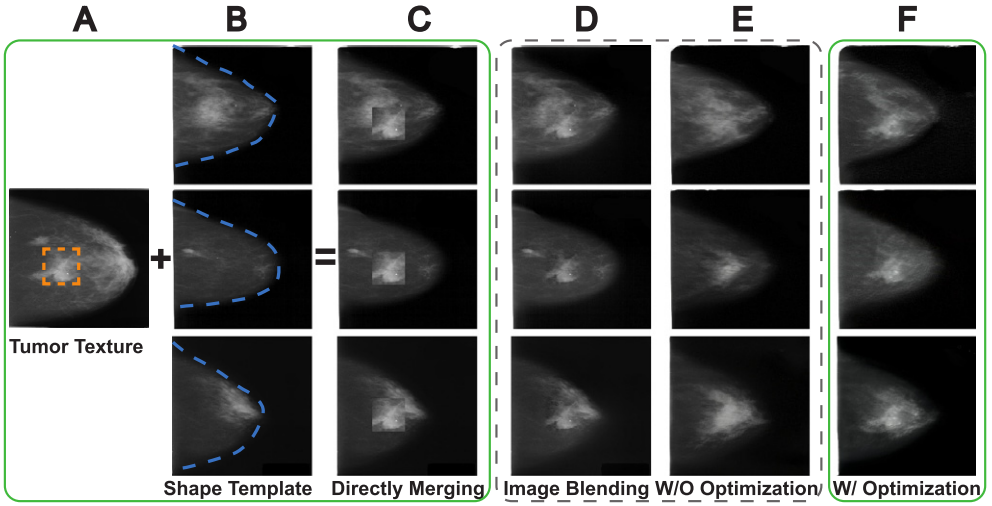

Figure 6. Attribute manipulation results. The desired image attributes are combined by merging the corresponding image patches (in Column A and B) directly. Then, the encoder will encode the manipulated image attributes, and the generator will produce the output correspondingly. After the final optimization, it is clear that the proposed method can generate the mammograms with the desired lesion texture and breast shape (Column F), compared to the results from the traditional image blending method (Column D) and the proposed method without the final optimization (Column E).

maintained. After the last step of optimization over the lesion patch, as it is shown in Fig. 6(F), the lesion texture is rendered. We also compare the results with the ones produced by a traditional image blending method. As it is shown in Fig. 6(D), the transition region between the lesion texture and the shape template background is not natural. Our proposed method can maintain both the breast shape and the lesion texture while generating authentic tissue texture.

4.4 *Human Evaluation*

To verify the authenticity of the generated images for different medical image modalities, we conducted an online psychophysical experiment, recruiting both untrained participants (i.e. no knowledge of medical imaging) and experts (e.g. radiologists or practicing clinicians who routinely read radiographs).

4.4.1 *Participants*

Six untrained observers (3 females, age range: 22–25) and seven experts (3 females, age range: 32–39) participated in the mammogram online survey. Two experts were excluded from the mammogram online survey (one dropped out and the other gave the same response on every trial). Five untrained observers (3 females, age range: 23–25) and seven experts (3 females, age range: 28–40) participated in the CT online survey.

All subjects reported to have normal or corrected-to-normal vision. Participants voluntarily participated and were offered $15 per hour as optional compensation. In our experience, radiologists typically refuse this modest compensation. The experiments were approved by the Institutional Review Board at the University of California, Berkeley. Participants provided informed consent.

4.4.2 *Stimuli*

For the mammogram online survey, 50 real mammograms and 50 fake (model generated) images were included. For the CT online survey, 50 real CT images and 50 fake CT images were presented. All the images were randomly selected from the corresponding data pools.

4.4.3 *Procedure*

The task was to rate each image from 0 (fake/generated image) to 10 (real image) in the data pool. Each individual image was shown for 5 s, and observers were asked to respond as quickly as possible. The experiment was self-paced, so observers viewed the stimuli as long as they wanted (up to 5 s), and they did not have time limit for giving responses. To ensure that participants did not randomly guess (or lapse), a small number of repetitive trials were also included in the online survey to establish a baseline test-retest reliability estimate. We compute the similarity among those repetitive trials.

4.4.4 *Results*

The results for mammogram and CT images in terms of the Receiver Operating Characteristic (ROC) curves are shown in Figure 7. For both untrained participants and radiologists, in the context of mammogram and CT images, their performance curves are near the diagonal (i.e. the chance level performance region), indicating that the generated medical images appeared authentic. The area under the curve (AUC) can also confirm the chance-level performance. The mean AUCs are 0.52 ($p = 0.395$, permutation test) and 0.60 ($p = 0.126$, permutation test) for untrained participants and radiologists respectively in mammogram online survey. The mean AUCs are 0.42 ($p = 0.888$, permutation test) and 0.42 ($p = 0.844$, permutation test) for untrained participants and radiologists respectively in CT online survey. As shown in the permutation tests, the large p-values indicate that performance is not statistically different from random performance.

Although the observers were not able to accurately discriminate real from fake images, this does not mean that observers randomly responded or failed to pay attention to the task. To confirm this, we calculated the test-retest reliability of each observers responses for repeated images. From the small number of repeated trials, the average test-retest similarity is 0.65, indicating "good" consistency. For a near-threshold task, the noise ceiling is not 1, and 0.65 is "good" in the sense that it is statistically reliable and significant [14, 20, 21]. The similarity is computed using Sokal-Michene metric [85]. It is noteworthy that observers can have high test-retest reliability despite low sensitivity (low AUC). The test-retest reliability indicates that observers tended to make the same judgments in repeated trials: they consistently confused some real (fake) images as being fake (real). This resulted in low sensitivity (low AUC) but consistent responses ("good" test-retest reliability).

We have appended the results of MRI and Skin Cancer images in the Appendix B to avoid redundancy. Results indicate that the generated medical images appeared authentic.

4.4.5 *Limitations*

Online studies have a range of potential limitations [4]. However, it has been well documented in the literature that online studies can reveal even very subtle psychophysical phenomena reliably, and these methods are now established [15, 62, 67]. In our online experiment, variations in the environment or monitor settings that might occur could add noise to the data, but they wouldn't generate the high test retest reliability we found, or the consistent pattern of results. The growing literature on internet-based psychophysics is consistent with this [67]. Moreover, we believe that our data adds a unique perspective on this issue: the advantages of online experiments are pronounced in cases where subjects are rare and/or very expensive to recruit, as is the case with the experienced and highly trained radiologist observers reported here. Future studies should consider online data collection for medical image perception tasks, in order to broaden representation, diversity, and improve sample sizes.

Another consideration with the experiments here is the images were viewed for a maximum of 5 s. The experiment was self-paced, and the participants could view the images as long as needed to make a choice, but this was limited to 5 s maximum viewing. There are both theoretical and empirical reasons that 5 s is likely to be sufficient for the task (see Appendix C), but it is conceivable that performance could change if observers were forced to view the images for prolonged periods of time. Future experiments should therefore examine the

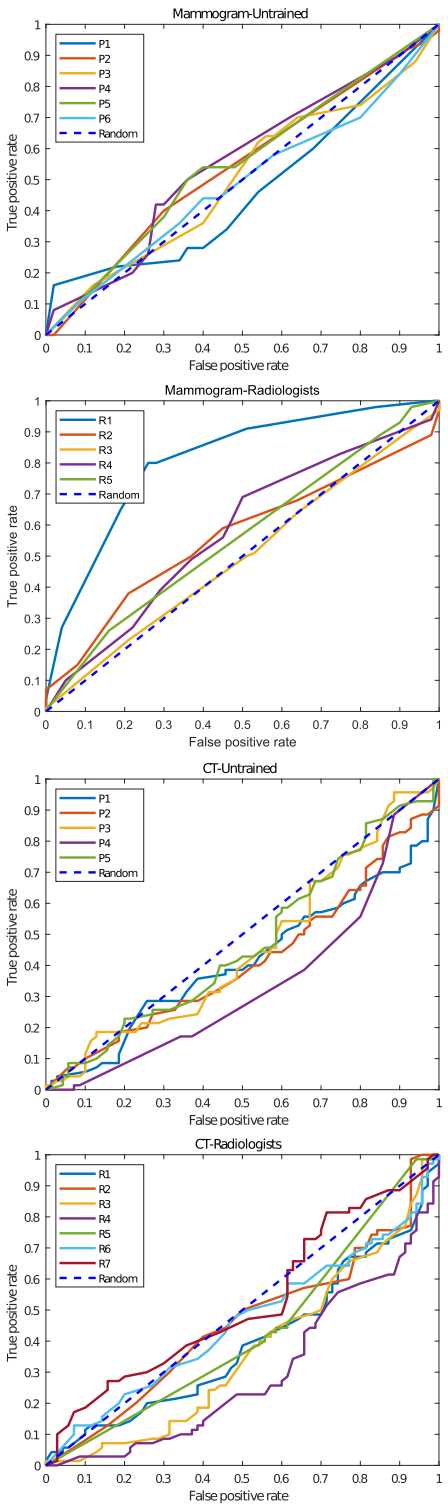

Figure 7. Human evaluation results. Participant performance is shown in the Receiver Operating Characteristic (ROC) curves. It is clear that their performance is near chance level (curves near the diagonal region), indicating that the generated medical images are authentic. Here, $P_1 - P_N$ and $R_1 - R_N$ represent different untrained observers and experts in corresponding experiments.

temporal integration of the visual processes that contribute to discrimination of near-metameric medical images.

Table I. Similarity Measurements for Mammogram Images.

	Image 1	Image 2
SSIM↑	**0.93**	0.89
PSNR (dB)↑	**38.99**	30.61
Perceptual↓	0.96	**0.64**

Table II. Similarity Measurements for MRI Images.

	Image 1	Image 2
SSIM↑	**0.84**	0.68
PSNR (dB)↑	**34.42**	33.01
Perceptual↓	3.89	**2.96**

Table III. Similarity Measurements for CT Images.

	Image 1	Image 2
SSIM↑	**0.54**	0.24
PSNR (dB)↑	**31.04**	29.91
Perceptual↓	27.77	**7.42**

4.5 *Perceptual Loss*

Currently, perceptual loss is utilized as a similarity metric in many computer vision tasks [37, 50, 87, 89]. In this section, we investigate the perceptual loss as a similarity metric in medical imaging domain. We compare its results with the results of Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), which are two common similarity metrics.

In the experiment, we utilize random samples from mammogram, MRI, CT, and skin cancer images as reference images (256×256). First, we apply traditional image distortions on those reference images, such as Gaussian blur, contrast distortion, geometric distortion, spatial shifting, and spatial rotation. Then, we calculate the similarity measurements for different outputs from traditional image distortions with respect to the reference images. Detailed computation algorithms can be found in Appendix A.

For quantitative comparison, we show the similarity measurement results in the following tables. For the SSIM and PSNR metrics, the larger the measurement is, the more similar it is between the measured image and the reference image (indicating by ↑). For perceptual metric, the smaller the measurement is, the more similar it is between the measured image and the reference image (indicating by ↓). Tables I, II, III and IV show the similarity measurements for mammogram, MRI, CT, and skin cancer images respectively.

For qualitative comparison, we compare the similarity measurement between Gaussian blur outputs (Figure 8 Image 1 Column) and the outputs from the rest of the traditional image distortions (Fig. 8 Image 2 Column). We

Table IV. Similarity Measurements for Skin Cancer Images.

	Image 1	Image 2
SSIM↑	**0.87**	0.74
PSNR (dB)↑	**36.26**	31.58
Perceptual↓	3.15	**1.99**

first asked human participants to give their choices of the image, which was more similar to the reference image. The results are labeled with green check marks as shown in Fig. 8. Then, according to the similarity measurements, we select the images which are preferred by SSIM/PSNR or perceptual loss metric. It is clear that SSIM and PSNR do not conform to human judgements. However, the similarity decisions from the perceptual loss metric are consistent with human judgements. Thus, the perceptual metric is more suitable for the similarity measurement in medical imaging area.

5. DISCUSSION

In this paper, we utilize Generative Adversarial Networks for medical image generation. Our results demonstrate generalizability of the proposed approach across different modalities, such as mammogram, MRI, CT, and skin cancer images. We also manipulate the generated images such that they contain desired attributes. Compared to traditional image blending methods which mainly edit locally, our proposed method not only embeds the desired image attributes but also edits the surrounding tissue texture accordingly to make the overall tissue texture distribution reasonable. Through adversarial training, the GAN model here learns an estimated manifold which is similar to the image manifold of the training dataset. This estimated manifold well-characterizes the semantic statistics of the training dataset, such as the tissue texture, tissue distribution, tissue shapes, and color distribution. Thus, once the contents of certain regions are altered, the GAN knows how to edit the surrounding region to match the semantic statistics of the training dataset, producing authentic manipulated images.

Our model can generate a vast range of possible stimuli that accomplish a range of specific and controllable goals. For example, the model can output specific body part shapes, lesion types and locations, background and tissue textures, etc. Additionally, our model is capable of generating morphed medical images, gradually and continuously. In certain medical image perception tasks, such as visual search [16, 82], visual detection and recognition [56], and decision making [75, 76], this kind of controllable medical image stimuli can be very useful. The intrinsic problem using real medical image data is that individual differences are substantial: it is not realistic to collect gradually morphing medical images from real medical image data (e.g., finding a sequence of naturally occurring tumors that smoothly morph between shapes or textures is highly unlikely). Using our proposed method, we can generate any number of authentic medical image stimuli that gradually morph. Moreover, all the images are generated via interpolation, which allows us to control the grain of the morphing.

For the perceptual loss metric, researchers [87] have determined that traditional similarity metrics, such as SSIM and PSNR, are not consistent with human perception of typical natural images. But deep neural network based perceptual metrics can, surprisingly, agree with human judgement. Through experiments, we arrive at the same conclusion in medical imaging domain as well; perceptual metrics preferred medical images and are more perceptually similar to the reference images compared to traditional similarity metrics. Thus, perceptual loss metric provides an important measurement of similarity in medical imaging. Using perceptual loss metric as the similarity measurement, we can also generate metamers for any specific medical image. The metamers are a cluster of perceptually similar images which have been widely used in perception researches.

Medical image perception research is a rapidly growing field. Typical approaches directly or indirectly assume that computer vision will be an alternative to clinical practice. Our study introduces an additional but very different perspective, which is to use computer vision to improve research on medical image perception. Clinicians will not be replaced anytime soon (if ever). To help clinicians make better judgments, we need to understand clinician perception, cognition, and decision. That requires having stimuli (datasets) that are simultaneously realistic (from the perspective of clinicians) and also controllable. Without this, it will be impossible to make the connection between the cognitive mechanisms that clinicians possess, and their diagnostic success in their practice.

Interestingly, the model and morphing approach we present here could be readily extended to three-dimensional volumetric images. Volumetric medical imaging is now gaining popularity as a standard practice in clinical setttings. The GAN model and morphing approach can be combined in future work to flexibly create volumetric data sets. Moreover, the GAN model is currently unconditioned. We can also change it to conditional GAN model such that changing certain part of the latent code (not through the encoder) can directly modify corresponding attributes of the output image.

6. CONCLUSION

In this paper, we propose usage of Generative Adversarial Network (GAN) for medical image generation. We tested our method on various medical image modalities such as mammogram, MRI, CT, and skin cancer images. Human evaluations verify the success of our method. We also adopt a controllable approach to manipulate the attributes of the generated images in order to meet certain experimental configurations. In the experiments, we successfully generate mammograms with the desired lesion texture and breast shape. The same approach can also be applied to MRI, CT, skin cancer images, and other medical imaging modalities. Finally, we compare traditional similarity measurements with the perceptual metric in medical imaging. We find that

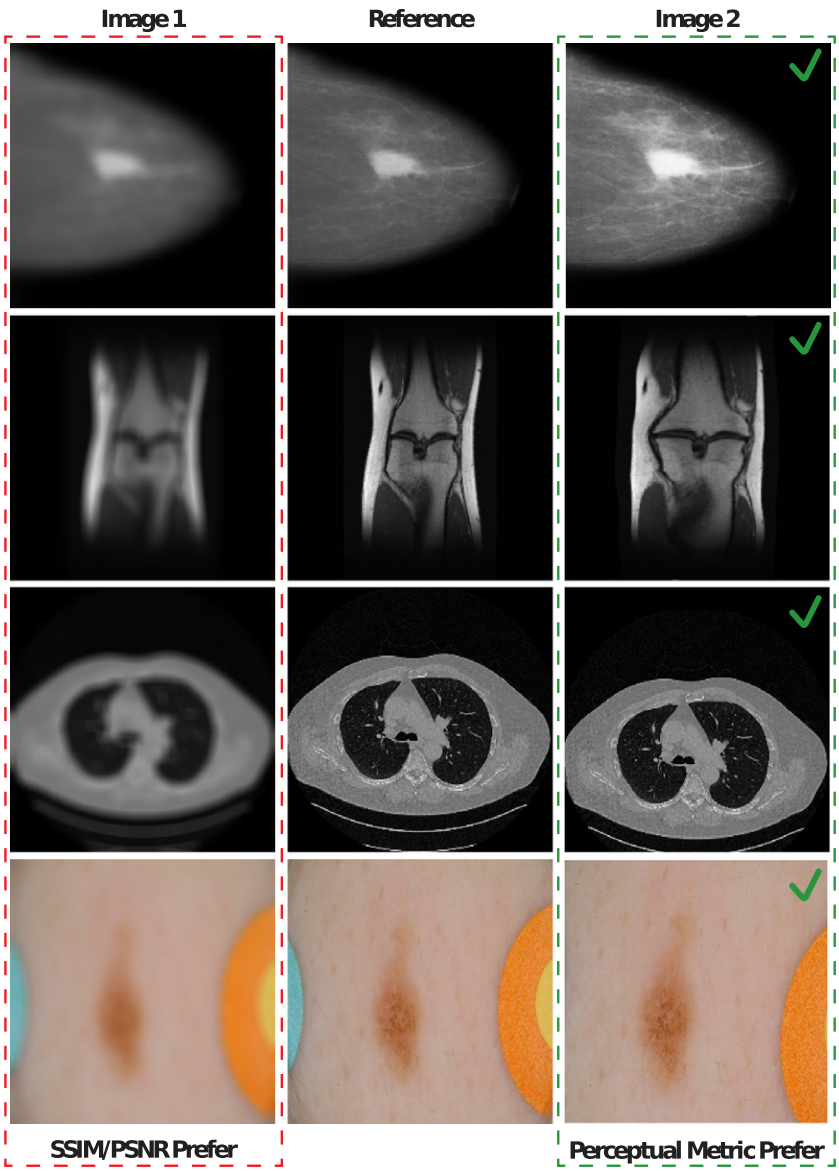

Figure 8. Which image is more similar to the reference? Image 1 Column shows the distortion by Gaussian blur. Image 2 Column shows the distortions by contrast distortion, geometric distortion, spatial shifting, and spatial rotation respectively. The human judgements are marked using green ticks. It is clear that SSIM/PSNR results do not conform to human judgements, while perceptual metric do.

the perceptual metric performs better than the traditional similarity metrics such as SSIM and PSNR.

APPENDIX A. SIMILARITY MEASUREMENTS

A.1 *SSIM*

The Structural Similarity Index Measure (SSIM) is computed over various patches of an image. The measure between two patches x and y of the same size is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{A1}$$

where μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator with $L = 2^{\#\text{bits per pixel}} - 1$, $k_1 = 0.01$, and $k_2 = 0.03$.

A.2 *PSNR*

Given a $m \times n$ reference image I and its distorted version K, the PSNR is defined as:

$$PSNR = 20\log_{10}(MAX_I) - 10\log_{10}(MSE), \tag{A2}$$

where MAX_I is 255 for 8-bit images, and the MSE is computed as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2. \tag{A3}$$

A.3 *Perceptual loss*

We utilize the same perceptual loss as [40]. The loss network is VGG-16 [71]. For the reference image r and the distorted image x, the perceptual loss is defined as:

$$L(x, r) = \lambda_c l^{\phi,j}_{\text{feat}}(x, r) + \lambda_s l^{\phi,J}_{\text{style}}(x, r), \tag{A4}$$

where λ_c and λ_s are scalars. In the experiment, we set $\lambda_c = 1$ and $\lambda_s = 1 \times 10^5$. ϕ represents the VGG network. $l^{\phi,j}_{\text{feat}}(x, r)$ is the feature reconstruction loss. Let $\phi_j(x)$ be the activation of the jth layer of the network ϕ with a shape of $C_j \times H_j \times W_j$. The feature reconstruction loss is defined as:

$$l^{\phi,j}_{\text{feat}}(x, r) = \frac{1}{C_j H_j W_j} \|\phi_j(x) - \phi_j(r)\|^2_2. \tag{A5}$$

The style reconstruction loss is defined as:

$$l^{\phi,J}_{\text{style}}(x, r) = \|G^{\phi}_j(x) - G^{\phi}_j(r)\|^2_F, \tag{A6}$$

where $G^{\phi}_j(x)$ is the Gram matrix with a shape of $C_j \times C_j$. The elements of the Gram matrix can be computed as:

$$G^{\phi}_j(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \tag{A7}$$

APPENDIX B. HUMAN EVALUATION OF GENERATED MRI AND SKIN CANCER IMAGES

We also collected human evaluation experiment data for MRI and Skin Cancer images. For the MRI experiment, four observers (1 expert, age range: 25–39) participated. For the Skin Cancer experiment, five observers (1 expert, age range: 20–39) participated. Unlike CT and mammogram image experiments, we could not recruit sufficient experts for MRI and Skin Cancer online surveys. All experiments were approved by the Institutional Review Board at UC Berkeley and the participants provided informed consent. Stimuli included 50 real and 50 corresponding fake images. All participants followed the same experimental procedures as described in Section 4.4.3.

The results for MRI and Skin Cancer images in terms of the Receiver Operating Characteristic (ROC) curves are shown in Figure B.1. The mean area under the curves (AUCs) are 0.57 ($p = 0.241$, permutation test) and 0.62 ($p = 0.123$, permutation test) for MRI and Skin Cancer respectively. Although we did not have experts for these MRI and Dermatology tests, we did have one trained radiologist participate and their data echoes that of untrained observers, all of which are consistent with the CT and mammogram data.

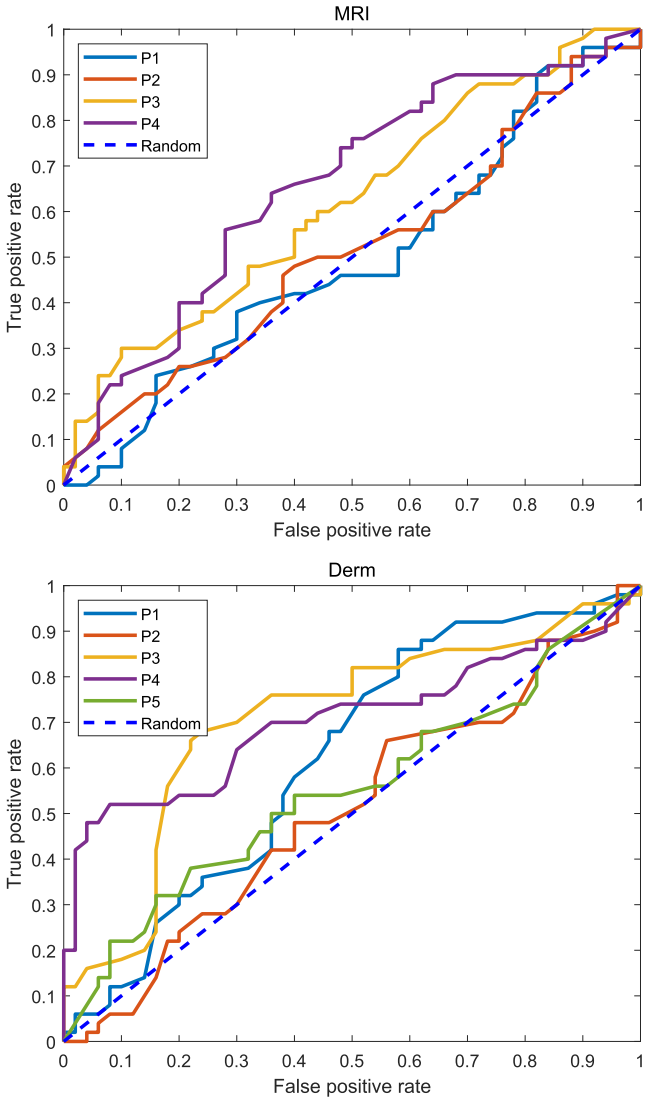

Figure B.1. Human evaluation results for MRI and Skin Cancer images. Participant performance is shown in the Receiver Operating Characteristic (ROC) curves. It is clear that their performance is near chance level (curves near the diagonal region), indicating that the generated medical images were authentic. Here, $P_1 - P_N$ represent different untrained observers and experts in corresponding experiments.

APPENDIX C. STIMULUS DURATION CONSIDERATIONS

There are both empirical and theoretical reasons for limiting the display to 5 s, and the empirical results confirm that 5 s was more than sufficient for observers to reach a reliable decision.

Firstly, previous research has demonstrated that radiologists can reliably discriminate radiographs within 1 s [11, 12, 18, 19, 24, 35, 39, 48, 55, 59]. In our experiment, we provide far more time than 1 s. Moreover, in self paced studies with static radiographs, radiologists often spend less than 5 s [69].

Secondly, our results show that accuracy does not vary with decision time. The decision time is reported as the time from the first viewing of the page to the final "submit" click by the observer. This is a conservative estimate of the decision duration. The relation between the error and decision time is shown in Figure C.1. The fitted line reveals that error and decision time are not correlated; more time did not make observers more accurate. Moreover, in this experiment, 60.0% of the decisions were made before stimuli

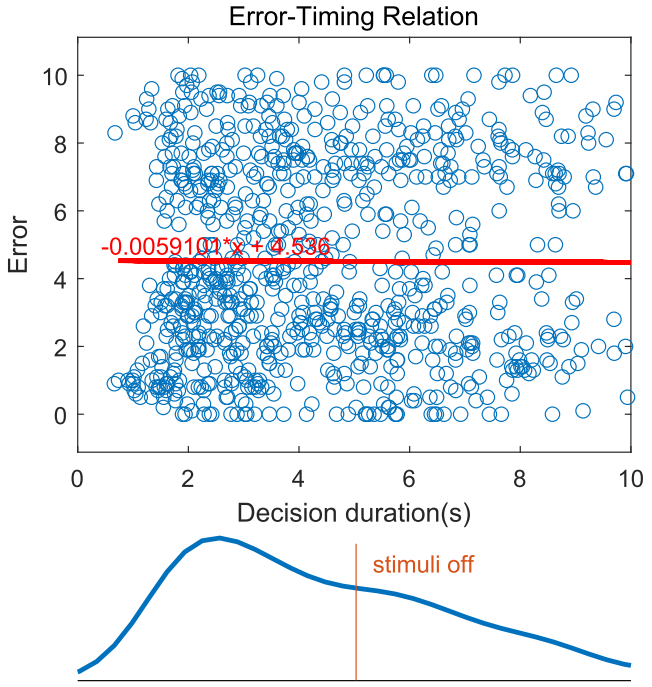

Figure C.1. Error-Timing Relation. The scatter plot shows the raw data of participants' error and their decision duration. We fit a linear function to reveal the relation between them. It is clear that the error and their decision time are not correlated. The bottom density distribution represents the distribution of participants' decision time. The orange line indicates the time point when stimuli disappeared. In this experiment, 60.0% of the decisions were made before the stimuli disappeared.

disappeared. Notably, the peak of response density does not occur at the 5 s boundary. It occurs around 2 s, which indicates that the 5 s stimulus duration limit does not induce pressure on participants' decision.

Finally, the significant test-retest reliability demonstrates that observers were consistent in their responses. If exposure duration limited performance, it would add noise and that test-retest reliability would be low [18].

Together, all of these considerations suggest that the duration of image interpretation was probably not the limiting factor. From the examples here, it appears that scrutinizing the real and generated images for more than a few seconds does not make them appear more or less similar. This hints that the metameric quality of the images is not due to a time constraint. Nevertheless, we did not force observers to scrutinize the images for more than 5 s, and it is conceivable that forcing an extended viewing of the images could improve performance. For this reason, it will be valuable in future studies to examine the temporal integration of the visual processes that contribute to discrimination of near-metameric medical images.

ACKNOWLEDGMENT

This work has been supported by National Institutes of Health (NIH) under grant # R01CA236793. We thank people who participated in the human evaluation and Min Zhou who recruited medical imaging experts from her hospital.

REFERENCES

[1] S. A. Amirshahi, M. Pedersen, and S. X. Yu, "Image quality assessment by comparing CNN features between images," J. Imaging Sci. Technol. **60**, 060410-1–060410-10 (2016).

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *Int'l. Conf. on Machine Learning* (PMLR, 2017), pp. 214–223.

[3] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba, "Seeing what a gan cannot generate," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2019), pp. 4502–4511.

[4] M. H. Birnbaum, "Human research and data collection via the internet," Annu. Rev. Psychol. **55**, 803–832 (2004).

[5] D. Bisla, A. Choromanska, R. S. Berman, J. A. Stein, and D. Polsky, "Towards automated melanoma detection with deep learning: Data purification and augmentation," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops* (IEEE, Piscataway, NJ, 2019), pp. 0–0.

[6] A. Bissoto, F. Perez, E. Valle, and S. Avila, "Skin lesion synthesis with generative adversarial networks," *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Orocedures, and Skin Image Analysis* (Springer, Cham, 2018), pp. 294–302.

[7] A. Bissoto, E. Valle, and S. Avila, "Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2021), pp. 1847–1856.

[8] K. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, M. Sallam, K. Chang, and K. Woods, "The digital database for screening mammography," *Third Int'l. Workshop on Digital Mammography* (Elsevier, Amsterdam, 1996), Vol. 58, p. 27.

[9] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *Int'l. Conf. on Learning Representations*. 2018.

[10] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, "Auto-gan: self-supervised collaborative learning for medical image synthesis," *Proc. AAAI Conf. on Artificial Intelligence* (AAAI, Palo Alto, CA, 2020), Vol. 34, pp. 10486–10493.

[11] D. P. Carmody, C. F. Nodine, and H. L. Kundel, "An analysis of perceptual and cognitive factors in radiographic interpretation," Perception **9**, 339–344 (1980).

[12] D. P. Carmody, C. F. Nodine, and H. L. Kundel, "Finding lung nodules with and without comparative visual scanning," Perception Psychophysics **29**, 594–598 (1981).

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems* (Morgan Kaufmann, San Francisco, CA, 2016), Vol. 29, pp. 2180–2188.

[14] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," Psychological Assess. **6**, 284 (1994).

[15] M. J. Crump, J. V. McDonnell, and T. M. Gureckis, "Evaluating amazon's mechanical turk as a tool for experimental behavioral research," PloS One **8**, e57410 (2013).

[16] T. Drew, K. Evans, M. L.-H. Võ, F. L. Jacobson, and J. M. Wolfe, "Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?," Radiographics **33**, 263–274 (2013).

[17] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," Preprint arXiv:1610.07629, (New York, NY, 2016).

[18] K. K. Evans, D. Georgian-Smith, R. Tambouret, R. L. Birdwell, and J. M. Wolfe, "The gist of the abnormal: Above-chance medical decision making in the blink of an eye," Psychonomic Bull. Rev. **20**, 1170–1175 (2013).

[19] K. K. Evans, T. M. Haygood, J. Cooper, A.-M. Culpan, and J. M. Wolfe, "A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast," Proc. Natl. Acad. Sci. **113**, 10292–10297 (2016).

[20] J. L. Fleiss, "Reliability of measurement," *The Design and Analysis of Clinical Experiments* (John Wiley & Sons, Hoboken, NJ, 1986).

[21] J. L. Fleiss, *Design and Analysis of Clinical Experiments* (John Wiley & Sons, Hoboken, NJ, 2011), Vol. 73.

[22] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," Neurocomputing **321**, 321–331 (2018).

[23] K. Fukushima, "Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron," IEICE Tech. Rep. A **62**, 658–665 (1979).

[24] A. Gale, J. Vernon, K. Miller, and B. Worthington, "Reporting in a flash," Br. J. Radiol. **63**, 71 (1990).

[25] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Proc. 28th Int'l. Conf. on Neural Information Processing Systems-Volume 1* (Morgan Kaufmann, San Francisco, CA, 2015), pp. 262–270.

[26] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, "Dermgan: Synthetic generation of clinical skin images with pathology," *Machine Learning for Health Workshop* (PMLR, Cambridge, MA, 2020), pp. 155–170.

[27] R. Girshick, "Fast r-cnn," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2015), pp. 1440–1448.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2014), pp. 580–587.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in Neural Information Processing Systems **27** (2014).

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," Advances in Neural Information Processing Systems **30** (2017).

[31] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "Gan-based synthetic brain mr image generation," *2018 IEEE 15th Int'l. Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, Piscataway, NJ, 2018), pp. 734–738.

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2017), pp. 2961–2969.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[34] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," J. Digit. Imaging **32**, 582–596 (2019).

[35] J. P. Houghton, B. R. Smoller, N. Leonard, M. R. Stevenson, and T. Dornan, "Diagnostic performance on briefly presented digital pathology images," J. Pathol. Inform. **6** (2015).

[36] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2017), pp. 1501–1510.

[37] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *Proc. European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2018), pp. 172–189.

[38] D. Hubel and T. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," J. Physiol. **148**, 574 (1959).

[39] T. Jaarsma, H. Jarodzka, M. Nap, J. J. van Merrienboer, and H. P. Boshuizen, "Expertise under the microscope: Processing histopathological slides," Med. Educ. **48**, 292–300 (2014).

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *European Conf. on Computer Vision* (Springer, Cham, 2016), pp. 694–711.

[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *Int'l. Conf. on Learning Representations* (2018).

[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 4401–4410.

[43] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 8110–8119.

[44] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," Preprint arXiv:1701.03056, (2017).

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Preprint arXiv:1412.6980, (2014).

[46] E. Kompaniez-Dunigan, C. K. Abbey, J. M. Boone, and M. A. Webster, "Adaptation and visual search in mammographic images," Attention Perception Psychophysics **77**, 1081–1087 (2015).

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems* (Morgan Kaufmann, San Francisco, CA, 2012), Vol. 25, pp. 1097–1105.

[48] H. L. Kundel and C. F. Nodine, "Interpreting chest radiographs without visual search," Radiology **116**, 527–532 (1975).

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2324 (1998).

[50] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 4681–4690.

[51] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal. **42**, 60–88 (2017).

[52] M. Manassi, Á. Kristjánsson, and D. Whitney, "Serial dependence in a simulated clinical visual search task," Sci. Rep. **9**, 1–10 (2019).

[53] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Int'l. Conf. on Artificial Neural Networks* (Springer, Berlin, Heidelberg, 2011), pp. 52–59.

[54] M. Mirza and S. Osindero, "Conditional generative adversarial nets," Preprint arXiv:1411.1784, (2014).

[55] M. D. Mugglestone, A. G. Gale, H. C. Cowley, and A. Wilson, "Diagnostic performance on briefly presented mammographic images," Proc. SPIE **2436** (1995).

[56] R. Nakashima, K. Kobayashi, E. Maeda, T. Yoshikawa, and K. Yokosawa, "Visual search of experts in medical image reading: the effect of training, target prevalence, and expert knowledge," Frontiers Psychol. **4**, 166 (2013).

[57] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," *Int'l. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer, Cham, 2017), pp. 417–425.

[58] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," *Int'l. Conf. on Machine Learning* (PMLR, Cambridge, MA, 2017), pp. 2642–2651.

[59] J. Oestmann, R. Greene, D. Kushner, P. Bourgouin, L. Linetsky, and H. Llewellyn, "Lung lesions: correlation between viewing time and detection," Radiology **166**, 451–453 (1988).

[60] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 2337–2346.

[61] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Preprint arXiv:1511.06434, (2015).

[62] S. Rajananda, M. A. Peters, H. Lau, and B. Odegaard, "Visual psychophysics on the web: open-access tools, experiments, and results using online platforms," J. Vis. **18**, 299–299 (2018).

[63] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," *2007 IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2007), pp. 1–8.

[64] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 779–788.

[65] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems **28** (2015).

[66] Z. Ren, S. X. Yu, and D. Whitney, "Controllable medical image generation via generative adversarial networks," *IS&T Electronic Imaging: Human Vision and Electronic Imaging* (IS&T, Springfield, VA, 2021), Vol. 2021, pp. 112–1–112–5.

[67] K. Semmelmann and S. Weigelt, "Online psychophysics: Reaction time effects in cognitive experiments," Behav. Res. Methods **49**, 1241–1260 (2017).

[68] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng. **19**, 221–248 (2017).

[69] H. Sheridan and E. M. Reingold, "The holistic processing account of visual expertise in medical image perception: A review," Frontiers Psychol. **8**, 1620 (2017).

[70] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," IEEE Trans. Med. Imaging **35**, 1285–1298 (2016).

[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Preprint arXiv:1409.1556, (2014).

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 1–9.

[73] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 2818–2826.

[74] M. Treviño, G. Birdsong, A. Carrigan, P. Choyke, T. Drew, M. Eckstein, A. Fernandez, B. D. Gallas, M. Giger, S. M. Hewitt, T. S. Horowitz, Y. V. Jiang, B. Kudrick, S. Martinez-Conde, S. Mitroff, L. Nebeling, J. Saltz, F. Samuelson, S. E. Seltzer, B. Shabestari, L. Shankar, E. Siegel, M. Tilkin, J. S. Trueblood, A. L. V. Dyke, A. M. Venkatesan, D. Whitney, and J. M. Wolfe, "Advancing research on medical image perception by strengthening multidisciplinary collaboration," JNCI Cancer Spectrum **6**, pkab099 (2022).

[75] J. S. Trueblood, Q. Eichbaum, A. C. Seegmiller, C. Stratton, P. O'Daniels, and W. R. Holmes, "Disentangling prevalence induced biases in medical image decision-making," Cognition **212**, 104713 (2021).

[76] J. S. Trueblood, W. R. Holmes, A. C. Seegmiller, J. Douds, M. Compton, E. Szentirmai, M. Woodruff, W. Huang, C. Stratton, and Q. Eichbaum, "The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making," Cogn. Res.: Princ. Implications **3**, 1–14 (2018).

[77] R. F. Wagner and D. G. Brown, "Unified snr analysis of medical imaging systems," Phys. Med. Biol. **30**, 489 (1985).

[78] S. Waite, A. Grigorian, R. G. Alexander, S. L. Macknik, M. Carrasco, D. J. Heeger, and S. Martinez-Conde, "Analysis of perceptual expertise in radiology–current knowledge and a new perspective," Frontiers Hum. Neurosci. **13**, 213 (2019).

[79] S. Waite, J. Scott, B. Gale, T. Fuchs, S. Kolla, and D. Reede, "Interpretive error in radiology," Am. J. Roentgenol. **208**, 739–749 (2017).

[80] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," Radiology **295**, 4–15 (2020).

[81] L. H. Williams and T. Drew, "What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures," Cogn. Res. Princ. Implications **4**, 1–24 (2019).

[82] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," Nature Hum. Behav. **1**, 1–8 (2017).

[83] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," J. Med. Imaging **5**, 036501 (2018).

[84] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, and M. Parente, "fastmri: An open dataset and benchmarks for accelerated mri," Preprint arXiv:1811.08839, (2018).

[85] B. Zhang and S. N. Srihari, "Properties of binary vector dissimilarity measures," *Proc. JCIS Int'l. Conf. Computer Vision, Pattern Recognition, and Image Processing* (Springer, Cham, 2003), Vol. 1.

[86] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," Med. Image Anal. **54**, 10–19 (2019).

[87] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2018), pp. 586–595.

[88] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain gan inversion for real image editing," *European Conf. on Computer Vision* (Springer, Cham, 2020), pp. 592–608.

[89] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," *European Conf. on Computer Vision* (Springer, Cham, 2016), pp. 597–613.