This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

medXGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space

Amil Dravid*[1], Florian Schiffers[1], Boqing Gong[2], Aggelos K. Katsaggelos[1]
[1]Northwestern University [2]Google
*amildravid2023@u.northwestern.edu

Abstract

*Despite the surge of deep learning in the past decade, some users are skeptical to deploy these models in practice due to their black-box nature. Specifically, in the medical space where there are severe potential repercussions, we need to develop methods to gain confidence in the models' decisions. To this end, we propose a novel medical imaging generative adversarial framework, medXGAN (**med**ical e**X**planation **GAN**), to visually explain what a medical classifier focuses on in its binary predictions. By encoding domain knowledge of medical images, we are able to disentangle anatomical structure and pathology, leading to fine-grained visualization through latent interpolation. Furthermore, we optimize the latent space such that interpolation explains how the features contribute to the classifier's output. Our method outperforms baselines such as Gradient-Weighted Class Activation Mapping (Grad-CAM) and Integrated Gradients in localization and explanatory ability. Additionally, a combination of the medXGAN with Integrated Gradients can yield explanations more robust to noise. The project page with code is available at: https://avdravid.github.io/medXGAN_page/.*

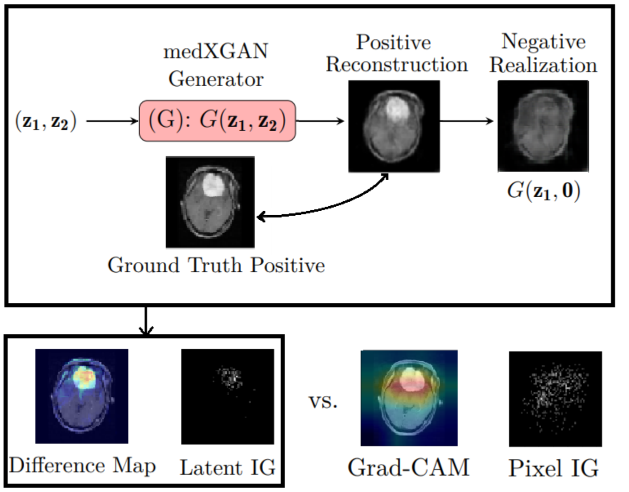


Figure 1. **Overview.** We propose a GAN framework medXGAN that takes in two latent vectors $(\mathbf{z1}, \mathbf{z2})$ to encode anatomical structure and classifier-specific pathology, respectively. After training with a fixed classifier providing feedback, the generator can be used to explain the classifier's decision. Given a ground truth positive image, the latent code can be found via an optimization scheme. The positive image then can be turned into a negative realization by relying on the medXGAN's latent sampling scheme. The classifier-specific features can be visualized via a pixel-wise difference between the negative and positive images or by integrating the gradients (IG) by traversing the latent space (LIG).

1. Introduction

Convolutional neural networks (CNNs) have enabled extremely accurate classification on large, complex datasets. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [41] kickstarted an era of massive efforts in tuning and finding new CNN architectures to beat classification benchmarks, among other tasks.

Despite their performance, neural networks are largely considered to be black boxes by the machine learning community [3]. As the size of these networks scale up with over millions of parameters [17], the black box becomes even more complex. Although they demonstrate strong performance on artificially set-up tasks on datasets such as ImageNet [41], neural networks have been found to be extremely sensitive. For instance, they perform poorly on data that is out-of-distribution with respect to their training set [40, 55]. Additionally, they break during inference on adversarial examples [15]. Adversarial examples are images from the distribution that have visually imperceptible perturbations but drastically change the classifier's output. These sensitivities drive the skepticism for deploying these models in actual practice.

Particularly, there are tremendous consequences in the medical domain. For example, with the onset of the COVID-19 pandemic, a slew of CNN models were created for COVID classification [4]. However, it has been found that many of them are trained on biased datasets, leading to a significant drop in performance on differently sourced datasets [9]. These models were misled by visualization and validation techniques such as Gradient-Weighted Class Activation Mapping (Grad-CAM) [44]. Following the surge of deep learning, the community has greatly increased efforts in explaining CNNs through various methods that we will explore later. However, many lack the ability to localize with fine detail [6] or have even been found to be model-agnostic and fail to key in on the most important features [1].

Generative Adversarial Networks (GANs) [14], a class of generative models, show promise in this task due to their ability to learn features and generate high fidelity images [22]. Additionally, incorporating domain knowledge of the underlying data into the visualization shows promise in creating higher quality explanations [18]. As such, we propose a novel GAN framework, medXGAN (**med**ical e**X**planation **GAN**), to visually explain what a medical image-based CNN classifier has learned. This substantially builds upon our prior work [11]. Our scheme relies on encoding domain knowledge of medical images into the generator's latent sampling scheme while incorporating a pretrained classifier into the original GAN formulation. Given an image of the target class, we can find the latent representation and interpolate with the image's negative realization to visualize changing class features according to the CNN.

Our contributions are as follows:

- We propose the medXGAN framework that uses a classifier to explicitly disentangle the latent code into anatomical structure and classifier-specific features. There is no need to search in the latent space for these corresponding factors.
- We encode domain knowledge of medical images into the latent sampling scheme using a continuous class code to obtain desirable latent interpolation properties.
- We propose using the negative realization of an image of the target class as a baseline for Integrated Gradients. We can then interpolate in the latent space, rather than pixel space, to obtain more localized and explanatory features.
- We demonstrate the promise of our method over baselines such as Grad-CAM and Integrated Gradients in localization ability and explanatory power using both quantitative and qualitative experiments.

2. Background

Generative Adversarial Networks (GANs) are a class of models that can generate new data from a target distribution [14]. A GAN consists of a generator network (*G*) and a discriminator network (*D*) that are typically parameterized as neural networks. Their training scheme is analogous to an art forger trying to fool an art appraiser. The generator takes in a *latent* or *noise* vector **z** drawn from a random prior distribution p_z, such as a spherical normal distribution. From this, it tries to create images $G(z)$ in order to fool the discriminator into classifying them as real. The discriminator takes turns looking at real images (x) from the true distribution (p_{data}) and generated images $G(z)$ and tries to classify them as real or fake correctly.

The GAN objective is grounded in game theory through a minimax game with :

$$\begin{aligned} \underset{G}{min}\ \underset{D}{max}\ & \mathop{\mathbb{E}}_{x\sim p_x}[\log D(x)] \\ & + \mathop{\mathbb{E}}_{z\sim p_z(z)}[\log(1-D(G(z)))] \end{aligned} \tag{1}$$

The generator seeks to minimize the Jensen-Shannon (JS) divergence between its estimated distribution p_g and the true distribution p_{data}. The generator is an *implicit density estimator*. It learns to sample from a distribution rather than explicitly parameterizing it. The discriminator tries to minimize the divergence between its distribution p_d and p_{data} and maximize the divergence of its estimated distribution for p_g with p_{data} [13]. Equilibrium occurs when $p_g = p_{data}$ and the discriminator's output is 0.5 for all images.

The conditional GAN (C-GAN) [32] is a natural extension, concatenating a discrete class code to the latent vector z to control the generator's ability to synthesize images from different categories. The Auxiliary-Classifier GAN (AC-GAN) builds upon this by formulating the discriminator to output an auxiliary classification for input images [35]. Works such as [27, 39] incorporate a separate classifier into the mix. Our work differs in that we include the classifier for the explicit objective of visualizing the classifier's learned attributes.

2.1. The GAN's Latent Space

The *latent space* refers to a low dimensional space that captures factors of variation of the data, such as angle, pose, lighting, etc [25]. The generator learns to sample from this manifold and produce high fidelity images. This is done by sampling a latent vector *z* drawn from some prior distribution. It has been found that interpolation and manipulation in this space can yield meaningful semantic results [38, 49].

Finding interpretable directions and learning representations that can separate informative factors of variations in the latent space is a highly active research topic [51]. The task of finding a latent space consisting of linear subspaces controlling factors of variation is known as *disentanglement* [22]. Various GAN-based approaches have found success in unsupervised, supervised, and semi-supervised regimes [7, 22, 28, 29, 33]. However, the corresponding factor for each subspace in these methods are arbitrary, and requires searching through them to find the factor of interest.

2.2. Visualization Methods

The two most common traditional visualization methods in medical image include Gradient-Weighted Class Activation Mapping (Grad-CAM) and Integrated Gradients [42, 44, 48]. As such, we will focus on these two.

Gradient-Weighted Class Activation Mapping (Grad-CAM). Gradient-weighted Class Activation Mapping (Grad-CAM) relies on the gradients of a target concept flowing into the final convolutional layer, resulting in a coarse localization map. This highlights the regions that maximally activate the CNN for a particular class [44]. This map is known as a *saliency map*.

First, through backpropagation, the gradient of the score y^c for class c is calculated before the softmax with respect to feature maps A^k of a convolutional layer k: $\frac{\partial y^c}{\partial A^k}$.

Next, these gradients are globally-averaged pooled to obtain α_k^c, which are neuron importance weights describing the importance of feature map k for a target class c. Lastly, a ReLU function is applied to a weighted combination of feature maps and their corresponding neuron importance weights to obtain a positive-influence saliency map:

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

A drawback of this method is its localization ability [6]. Furthermore, it depends on the size of the convolutions. So, it tends to be biased towards larger models. In the medical domain, the lack of fine-grained detail can inadvertently capture a disease feature by the nature of ”casting a wide net,” leading to false confidence [50]. Additionally, this saliency map cannot tell the ”whole story” and explain how the predicted features contribute to the prediction [43].

Integrated Gradients (IG). Integrated Gradients (IG) relies on attributing the prediction of a deep network to the pixels of its input image [48]. Given a target image x to visualize, a baseline image x' is also established. There are many choices, but a completely black image is common [30]. However, choosing an appropriate baseline image is an open problem [47]. From there, a pixel-wise interpolation between these two images is fed into the classifier f. The gradient is then taken with respect to the input pixels. The parameter α governs the scale of interpolation.

As the interpolation from the black image approaches the target image, the gradients are accumulated and averaged. This leads to a map that highlights pixels that contain negative or positive attribution to the target class. This is formulated as:

$$\mathbf{IG} = (x - x') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x} d\alpha \tag{3}$$

although it is discretized with summations in practice.

Despite its ability to attribute importance at the pixel level rather than patch level as Grad-CAM does, Integrated Gradients is highly dependent on the chosen baseline [47]. Additionally, it can pick up noise and amount to an edge detector [1, 10].

Generative-Based Visualization. Generative models have been proposed to visualize classifiers [26, 31, 36, 43]. The work in [36] relies on Variational Autoencoders [24], but is limited by experiments on artificial toy datasets. The methods in [26, 43] rely on StyleGAN [22] that generate high quality explanations, but lack substantial quantitative experiments on common baselines, and rely on search algorithms to find the relevant latent codes.

Our work is able to explicitly disentangle the latent code in a highly structured manner. Additionally, we show qualitative and quantitative experiments over the common baselines such as Grad-CAM and Integrated Gradients. The latent space in medXGAN is also optimized for meaningful latent interpolation that leads to the extension of Integrated Gradients in the latent space.

3. Methods

Utilitizing the medXGAN for feature visualization consists of three steps. First the classifier must be pretrained, and then incorporated into the training of the medXGAN. Then, given a ground truth positive image, a reconstruction task enables discovery of the latent vectors. Lastly, the latent code can be used to generate a negative realization of the positive image. This yields powerful visualization capability as we can traverse the latent space to observe changing features, among other visualization methods.

3.1. medXGAN Overview

In order to visualize a CNN, we incorporate the pretrained network (C) into the original GAN framework (see Fig. 2). The weights of the generator (G) and discriminator (D) are trained, playing the typical minimax game with real samples x and generated samples $G(\cdot)$. The weights for the classifier are fixed, thus this network provides feedback to the generator on the class (y) according to the CNN's learned distribution p_c . The overall objective is:

$$\begin{aligned} \underset{G}{min}\, \underset{D}{max} & \underset{x \sim p_x}{\mathbb{E}} [\log D(x)] \\ & + \underset{z_1 \sim p_{z_1}, y \sim p_y}{\mathbb{E}} [\log(1 - D(G(z_1, y)))] \\ & - \underset{z_1 \sim p_{z_1}, y \sim p_y}{\mathbb{E}} [\log(p_c(y|G(z_1, y)))] \end{aligned} \tag{4}$$

where the first two terms correspond to the original GAN formulation, and the third term relates to incorporating class features according to the CNN. The generator takes in two latent vectors that are concatenated. z_1 is drawn from a

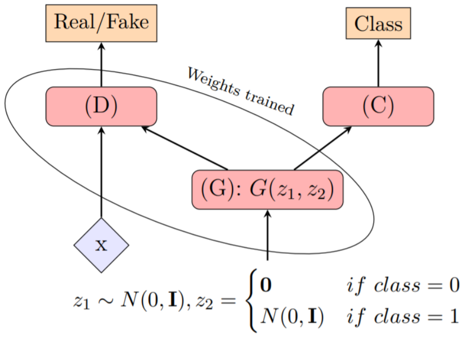

Real/Fake
Class
Weights trained
(D)
(C)
(G): $G(z_1, z_2)$
x
$z_1 \sim N(0,\mathbf{I}), z_2 = \begin{cases} \mathbf{0} & if\ class = 0 \\ N(0,\mathbf{I}) & if\ class = 1 \end{cases}$

Figure 2. **medXGAN training scheme.** A pretrained classifier provides class feedback to the generator's synthesized image while the discriminator and generator play their typical adversarial game. The latent vector consists of $\mathbf{z_1}$ which encodes anatomical structure and $\mathbf{z_2}$ which corresponds to a continuous class code for pathology features according to the classifier.

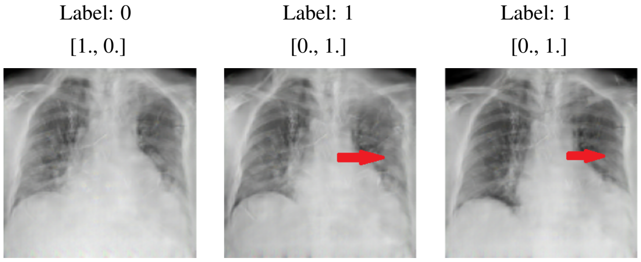


Figure 3. **Example of disentangled lungs and classifier features.** The classifier's softmax outputs are above along with label provided to generator. We can see that the lung and skeletal structure is intact, but features within the anatomy change, leading to different classifier outputs. The largest changes are highlighted by an arrow.

spherical normal distribution and corresponds to anatomical structure. z_2 corresponds to pathology features according to the classifier. If the image to be generated is negative (absent pathology) then z_2 is assumed to be $\mathbf{0}$. Otherwise, it is drawn from a spherical normal distribution.

After training the GAN, we can now visualize the CNN. Given a real positive image, we can find its latent representation via stochastic gradient descent (SGD):

$$\arg\min_{z_1,z_2} \text{MSE}(G(z_1,z_2),x) + \text{BCE}(C(G(z_1,z_2)),C(x)) \tag{5}$$

where we are trying to match the pixels of the true image and reconstruction via a pixel-wise mean-squared error (MSE). We also match the classifier's output for both images with a binary cross-entropy loss(BCE). After we find z_1 and z_2, we can rely on the sampling scheme for z_2, changing it to $\mathbf{0}$ in order to convert the positive image to a negative realization with high confidence while retaining the same anatomical structure.

Finally, we can interpolate in the latent space between the negative and positive images to visualize how the classifier's output changes with the interior pathology. We interpolate through the latent vector z_2 with steps n at a rate of λ while keeping the anatomical structure constant with z_1 by looking at the outputs of $G(z_1, \mathbf{0} + n\lambda z_2)$, for $n = 1, 2, ...$

3.2. A Disentanglement Perspective

Mutual information describes the amount of information obtained about one random variable by observing the other random variable. Given two random variables, (x, y), mutual information I is related to entropy H: $I(x;y) = H(x) - H(x|y) = H(y) - H(y|x)$. It has been found that maximizing mutual information between some feature code y and image x can lead to disentangled representations [7]. We can find a variational lower bound [2] for the mutual information $I(y; G(z_1, y))$. This method relies on the fact that the KL divergence between the posterior of the classifier's learned distribution and the true posterior is non-negative.

$$\begin{aligned} I(y;G(z_1,y)) &= H(y) - H(y|G(z_1,y)) \\ &\geq H(y) - \mathop{\mathbb{E}}_{z_1\sim p_{z_1}, y\sim p_y}[\log(p_c(y|G(z_1,y)))] \end{aligned} \tag{6}$$

$H(y)$ can considered a constant term, so maximizing the mutual information between the class code and the generated image amounts to minimizing $\mathbb{E}[\log(p_c(y|G(z_1,y)))]$, which corresponds exactly to the third term of Eq. 4, thus leading to disentangled representations. We can see an example of this in Fig. 3.

3.3. A Manifold Perspective

According to the *manifold hypothesis*, high dimensional data lies on lower dimensional manifolds in this space [12]. However, natural data lies on a union of disjoint manifolds, and GANs struggle to model a distribution supported on disconnected manifolds [23]. Interpolating between samples on disjoint manifolds may result in off-manifold or unnatural samples. The Conditional-GAN induces disconnectedness by using a discrete code. In our case, we want ”semantically smooth” interpolation in the latent space, with the classifier's output monotonically increasing as we interpolate from a negative and positive realization of a medical image. This lends itself to a smooth Integrated Gradients visualization that does not pick up on spurious features. As such we propose a continuous code that encodes domain knowledge of medical images. Typically there is an underlying anatomical structure that is fixed, but the disease pathology is not deterministic, and can manifest in multiple ways within the anatomy. As such, there is one realization of the negative image with $z_2 = \mathbf{0}$, and multiple realizations of a positive image with $z_2 \sim \mathcal{N}(0, \mathbf{I})$.

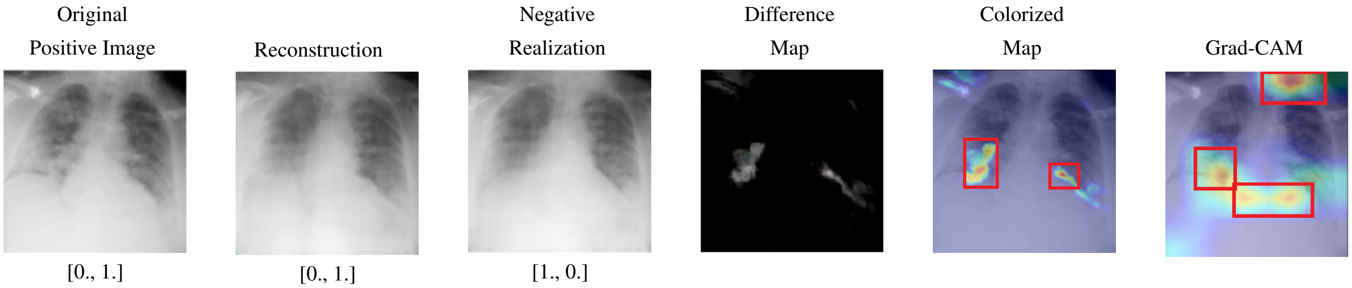


Figure 4. **Reconstructing a COVID positive image and turning it negative**. Given a real positive class image, the latent code $(\mathbf{z_1}, \mathbf{z_2})$ can be found via SGD with the generator to match pixel-wise and through the classifier's output. The positive reconstruction can be turned into a negative realization by setting $\mathbf{z_2} = \mathbf{0}$. The classifier softmax outputs are below the respective images. Here, we visualize the pixel-wise difference between the realizations. Compared to Grad-CAM, we see more localization ability.

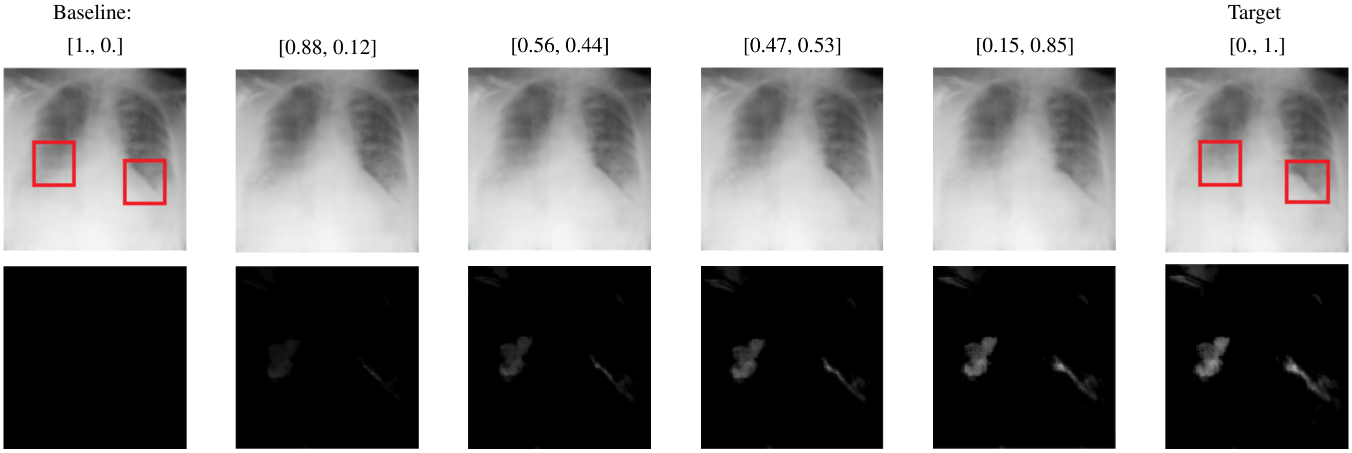


Figure 5. **Example of latent interpolation** Because $\mathbf{z_1}$ and $\mathbf{z_2}$ are disentangled latent codes, we can traverse the latent space by fixing $\mathbf{z_1}$ and stepping through $\mathbf{z_2}$. We observe fixed anatomical structure and changing pathology according to the classifier. The classifier's softmax outputs shown above each respective image. Additionally, we visualize the accumulating absolute value pixel-wise difference through the interpolation, which illustrates how the changing features contribute to the classifier's output.

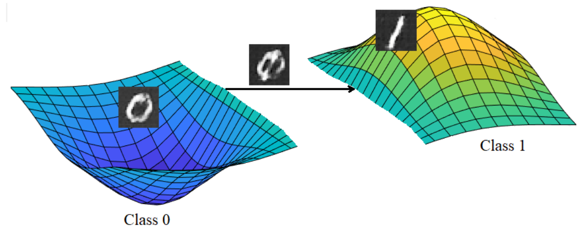


Figure 6. **Disconnected manifolds.** Interpolating between disconnected manifolds can lead to off-manifold intermediate results, which does not lend itself to a smooth and meaningful latent interpolation. For instance, the intermediate result for interpolating between 0 and 1 does not make natural sense. With medical images, we want a smooth transition with clinically plausible images.

4. Experiments

We present qualitative analysis as well as quantitative experiments of our medXGAN method against the most popular explanatory techniques of Grad-CAM and Integrated Gradients. To begin, we first measure how well the generator captures the classifier's distribution. We first trained a VGG-16 network [46] to classify COVID-19 on an in-house dataset of COVID chest X-rays resized to 128×128 [54]. This network achieves roughly $75.0\% \pm 0.9$ accuracy on the dataset, which was upsampled to become class balanced. Additionally, we trained an off-the-shelf CNN for binary classification of the presence of various brain tumor types on 64×64 MRIs [5], achieving $85.20\% \pm 1.3$ accuracy. The area under the receiver-operator (AUC) score for these is roughly of the same magnitude as the accuracy.

We generated 4 images using the same anatomical structure by keeping z_1 fixed, with 1 negative and 3 positive realizations using the z_2 sampling scheme. This was repeated 1000 times to generate 4000 total images which were fed into the respective classifier for classification. For the MRI dataset, the classifier correctly predicts the class given to the generator with accuracy $\mathbf{99.2}\% \pm 0.2$. For

Data		Accuracy	AUC
MRI	Generated	**99.2%** ± 0.2	**0.995** ± 0.001
	Real	85.2% ± 1.3	0.975 ± 0.005
X-Ray	Generated	**93.7%** ± 0.3	**0.980** ± 0.003
	Real	75.0% ± 0.9	0.763 ± 0.008

Table 1. **Accuracy and AUC of classifier for tumor and COVID-19 classification on true validation images and generated data.** The high accuracy on the generated data suggests that the medXGAN generator is fitting strongly to the classifier's learned distribution based on its training data.

the COVID dataset, the accuracy is **93.7**% ± 0.3. Thus, there is a strong correspondence between the generator and classifier's distributions as the generator is incorporating classifier-specific features.

4.1. Grad-CAM Experiment

Many experiments in the explainable AI space rely on counterfactual reasoning: "what would happen if we changed this feature?" Along these lines, we use Grad-CAM to create saliency maps for both brain MRIs and chest X-rays. Additionally, we use medXGAN to reconstruct negative and positive realization of the images, and take a pixel-wise difference to highlight the important changing features. After localizing the features through the two methods, we perturb the salient features and observe the change in the classifier's output on these new images, a commonly employed metric [36, 37]. Although we can use perturbations such as Gaussian noise, or black or white pixels, we opt to replace the salient pixels with the average intensity of the image, and observe the average drop in the classifier's "positive" softmax output for multiple images. Given the grayscale images, black or white pixels may bias the decision of the model towards a particular class. Additionally, the output would be sensitive to the particular instance of random noise. For a fair evaluation with Grad-CAM, we do the "averaging perturbation" for the medXGAN features as well instead of taking the negative realization of the positive image. The results of the counterfactual experiments are summarized in Tab. 2, which indicates that the medXGAN is able to identify features with more explanatory power due to the greater drop in the classifier's output.

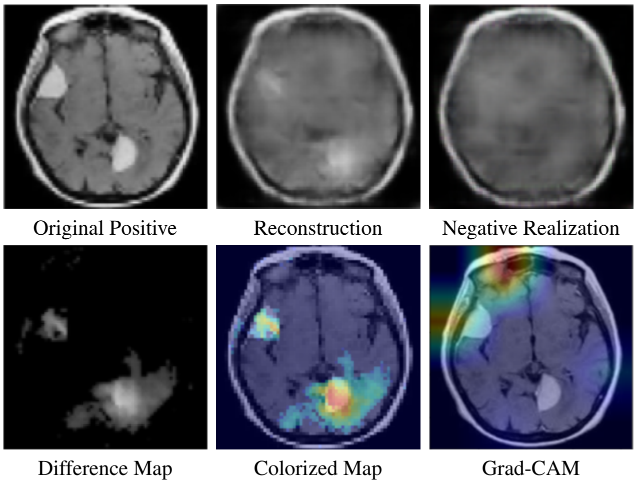
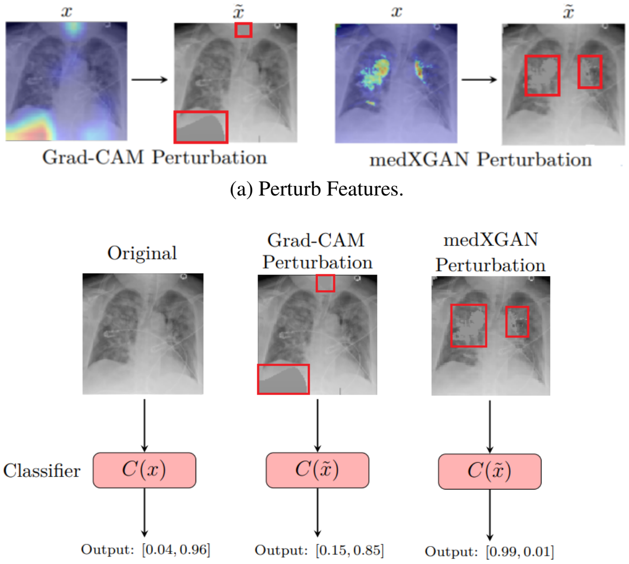


Figure 7. **Example of Grad-CAM and medXGAN methods on Brain MRIs for feature visualization.** In this example, despite lacking a perfect reconstruction, the medXGAN method localizes two tumors, while Grad-CAM focuses on one tumor and an eye.



(a) Perturb Features.

(b) Input into classifier.

Figure 8. **Example of counterfactual perturbations**. We first feed a positive image from the validation set into a classifier and observe the classifier's output. Then, given the features highlighted from medXGAN and GradCAM, we perturb them to the mean intensity value of the original image. We then feed them into the classifier to observe the drop in confidence for the positive class. In this example, we see a larger drop in the classifier's output from the medXGAN method.

4.2. Integrated Gradients Experiment

As visualizing brain tumors is more interpretable to non-experts, we opt to use Integrated Gradients with just the brain MRI dataset. For this experiment, we measure the degree of localization. We apply the standard Integrated Gradients to MRIs with tumors by interpolating in the pixel space between a black image and the target image as is standard practice. Additionally, we propose to use our medXGAN to interpolate in the latent space between the negative and positive realizations of the image, which is one of our novel contributions. We refer to this as Latent Integrated

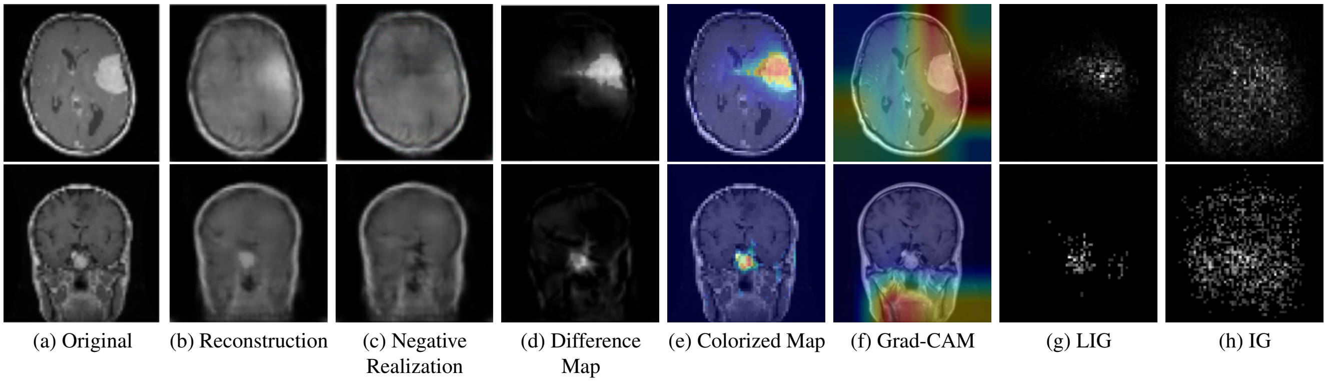


Figure 9. **Example of visualizations on the brain MRI dataset.** Both difference maps and integrating gradients through the latent space (LIG) localize important classifier-specific features much finer than Grad-CAM and standard pixel-wise Integrated Gradients (IG).

	MRI	X-Ray
medXGAN	**0.97** ± 0.03	**0.91** ± 0.08
Grad-CAM	0.89 ± 0.10	0.83 ± 0.15

Table 2. **Results for counterfactual experiments**. We perturb features identified by medXGAN and Grad-CAM by changing them to the average intensity of the image. After feeding the image into the classifier, we observe the average drop in softmax output for positive. The medXGAN sees a larger drop in classifier confidence, suggesting that its identified features correspond more strongly to the classifier's decisions.

Gradients (LIG). This can be formulated as:

$$\mathbf{LIG} = [G(z_1, z_2) - G(z_1, \mathbf{0})] \int_{\alpha=0}^{1} \frac{\partial f(G(z_1, \mathbf{0} + \alpha z_2))}{\partial x} d\alpha \quad (7)$$

where were are taking the gradient with respect to the input x to the model $f(x)$ as we interpolate in the latent space. Afterwards, we find the ratio of pixels with some non-zero attribution value through latent vs. pixel interpolation. The averaged ratio over multiple images was $\mathbf{0.21} \pm .09$, indicating that our method is able to localize the salient features with much finer detail. Essentially, it is using one-fifth the number of pixels that standard Integrated Gradients attributes. Our method does not capture as much noise or as many spurious edges as the baseline does (Fig. 9).

4.3. Qualitative Analysis

In both the brain tumor and COVID dataset, we see that the medXGAN is able to attribute classifier-specific features that are more fine-grained compared to Grad-CAM or Integrated Gradients. For instance, medXGAN is able to completely capture the brain tumor in Fig. 9 while Grad-CAM misses it. Although the reconstructed images are not of the highest fidelity compared to the ground truth, they still capture the important anatomical structure and pathological features, lending to successful visualization. Additionally, we notice that through latent interpolation, the classifier's output for positive is monotonically increasing, a result based on connecting two disconnected class manifolds through a continuous class code. While traditional visualization methods will give a single map highlighting salient regions, our method can be employed to study how the features contribute to the classifier's output through latent interpolation. Additionally, we implicitly get access to the classifier's decision boundary as we can observe when the classifier changes predictions based on the latent interpolation (Fig. 5).

5. Convergence Experiments

As visualization with the medXGAN relies on finding the latent code of ground truth images, we quantitatively examine how visualizations can change depending on different runs of the optimization scheme. We randomly initialize the latent vectors with values from a standard normal distribution. For the medxGAN trained on the brain MRIs, the latent vectors are $z_1 \in \mathbb{R}^{1000\times1}$ and $z_2 \in \mathbb{R}^{100\times1}$. For the GAN trained on chest X-rays, they are $z_1 \in \mathbb{R}^{100\times1}$ and $z_2 \in \mathbb{R}^{10\times1}$. On various positive class images from the MRI and X-ray datasets, we run stochastic gradient descent on $(\mathbf{z_1}, \mathbf{z_2})$ multiple times for a fixed $10,000$ epochs. We then compute two metrics. First, we find the pairwise cosine similarity between the latent vectors found between the multiple runs. This measures the similarity in the latent space. We then run the vectors through the generator and measure the perceptual similarity via the Structural Similarity Index Measure (SSIM) [52]. These results are summarized in Tab. 3. As these scores all tend towards 1, we see that despite the nonconvex optimization scheme, we are converging to very similar regions in the image and latent space. As the cosine similarity and SSIM are both approaching one, it appears that convergence in the latent

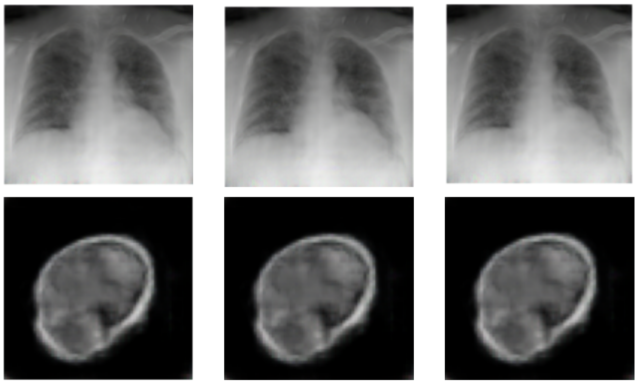


Figure 10. **Example of converging reconstructions.** Given random initial seeds for the latent vectors, the reconstructions appear to converge to extremely similar images among each other.

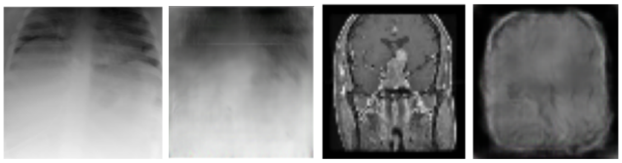
space correlates to convergence in the image space. However, this can be further studied.

	Cosine Similarity	SSIM
MRI	0.965 ± 0.008	0.970 ± 0.020
X-Ray	0.994 ± 0.005	0.968 ± 0.015

Table 3. **Convergence results.** Based on random initial seeds, we optimize over the latent vectors for 10,000 epochs to reconstruct various images. We compute the SSIM between the respective images as well as the cosine similarity between the latent vectors. With the metrics approaching 1, the reconstructions seem to converge in the image and latent space.

6. Limitations and Discussion

Despite the promising results of medXGAN, we recognize some limitations of our method. One is based on training data. It is well-known that GANs require a significant amount of training data in order to faithfully ”learn” the training distribution [34]. In our experiments, the generation of chest x-rays are much higher fidelity than the brain MRIs due to the dataset sizes: $\sim$ 15000 vs. $\sim$ 2000 images. Nevertheless, even with the brain MRIs, we are able to capture important classifier-specific features. To extend the generator's capacity given limited data, we suggest employing a data augmentation scheme such as [21] or a transfer learning approach like [56]. Additionally, faithful image reconstruction relies on the GAN learning a rich latent space. If the generator becomes too adapted to the training distribution, examples out-of-distribution or ”off-manifold” may result in poor reconstructions [20, 53]. In Fig. 11, we see examples of poor reconstructions as the ground truth images are not well-represented in the distribution. We plan to scale up the medXGAN to higher resolution and more optimized frameworks for the highest fidelity image synthesis. However, we have seen that even rough approximations of the ground truth images can be very powerful. Additionally, given the restriction to binary classification, our next steps include extending to multi-class classification.



Original X-Ray Reconstruction Original MRI Reconstruction

Figure 11. **Failed reconstructions.** These original images can be considered anomalies as some of the characteristics they present are not well represented in the dataset. For instance both the X-ray and MRI are heavily zoomed in. Although the reconstruction captures some structure, it misses many details within.

Despite the growing field of explainable AI, quantitative benchmarking and comparison is still an open problem [8]. Evaluation can largely be *ad hoc*. For instance, we determined that using the mean intensity of the image, would serve as the fairest perturbation without inducing bias [19]. We could also use an image inpainting technique instead [16]. We can also further validate our method by measuring its reliance on the classifier's weights as well as the dataset labels through methods proposed in [1].

Despite the computational overhead of training a GAN for visualization, the generator can also provide meaningful data augmentation to further optimize the classifier [45]. With the disentangled latent codes, users have more control over the generated images. Additionally, the latent space is optimized so that interpolation leads to a monotonic increase in the classifier's output. As such, this can be leveraged to create samples near the classifier's decision boundary.

7. Conclusion

In this work, we presented medXGAN, a novel GAN framework that encodes domain knowledge of medical images into its latent sampling scheme through a continuous class code. This allows for explicit disentanglement of anatomical structure and classifier-specific pathology features. Additionally, we proposed using a negative realization of a positive class image as a baseline along with latent interpolation for Integrated Gradients. We establish this as Latent Integrated Gradients (LIG). We also demonstrated medXGAN's promising explanatory and localization power through quantitative and qualitative analysis over the baselines of Grad-CAM, Integrated Gradients. It is important to note that the visualizations are not what the actual class features should be, but rather what the classifier thinks. So the visualizations are subject to the biases and errors of the classifier. Ultimately, we hope our method inspires further efforts to open the black box of neural networks.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 2, 3, 8

[2] David Barber Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004. 4

[3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 1

[4] Zaid Abdi Alkareem Alyasseri, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A Awadallah, Ammar Kamal Abasi, Sharif Naser Makhadmeh, Osama Ahmad Alomari, Karrar Hameed Abdulkareem, Afzan Adam, Robertas Damasevicius, et al. Review on covid-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems*, page e12759, 2021. 2

[5] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri), 2020. 5

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 2, 3

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 2, 4

[8] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020. 8

[9] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. 2

[10] David Drakard, Rosanne Liu, and Jason Yosinski. Exploring unfairness in integrated gradients based attribution methods. 2021. 3

[11] Amil Dravid and Aggelos K Katsaggelos. Visual explanations for convolutional neural networks via latent traversal of generative adversarial networks. *arXiv e-prints*, pages arXiv–2111, 2021. 2

[12] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 4

[13] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[16] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2013. 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[18] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021. 2

[19] Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. In *International Conference on Learning Representations*, 2021. 8

[20] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021. 8

[21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 8

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3

[23] Mahyar Khayatkhoei, Maneesh K Singh, and Ahmed Elgammal. Disconnected manifold learning for generative adversarial networks. *Advances in Neural Information Processing Systems*, 31, 2018. 4

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[25] Line Kuhnel, Tom Fletcher, Sarang Joshi, and Stefan Sommer. Latent space non-linear statistics. *arXiv preprint arXiv:1805.07632*, 2018. 2

[26] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 3

[27] Chongxuan Li, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *Advances in neural information processing systems*, 30, 2017. 2

[28] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International Conference on Machine Learning*, pages 6127–6139. PMLR, 2020. 2

[29] Bingchen Liu, Yizhe Zhu, Zuohui Fu, Gerard De Melo, and Ahmed Elgammal. Oogan: Disentangling gan with one-hot sampling and orthogonal regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4836–4843, 2020. 2

[30] Daniel Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method

and extensions to internal neuron attributions. *arXiv preprint arXiv:2202.11912*, 2022. 3

[31] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual-counterfactual explanations for medical non-experts using generative adversarial learning. *arXiv preprint arXiv:2012.11905*, 2020. 3

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[33] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar. Semi-supervised stylegan for disentanglement learning. In *International Conference on Machine Learning*, pages 7360–7369. PMLR, 2020. 2

[34] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 8

[35] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017. 2

[36] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 33:5453–5467, 2020. 3, 6

[37] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2304, 2021. 6

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016. 2

[39] Harsh Rangwani, Konda Reddy Mopuri, and R Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Uncertainty in Artificial Intelligence*, pages 1618–1627. PMLR, 2021. 2

[40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[42] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *medRxiv*, pages 2021–02, 2021. 3

[43] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021. 3

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 3

[45] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 8

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[47] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. 3

[48] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[49] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020. 2

[50] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 3

[51] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514–63537, 2020. 2

[52] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[53] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2019. 8

[54] Ramsey M Wehbe, Jiayue Sheng, Shinjan Dutta, Siyuan Chai, Amil Dravid, Semih Barutcu, Yunan Wu, Donald R Cantrell, Nicholas Xiao, Bradley D Allen, et al. Deepcovid-xr: an artificial intelligence algorithm to detect covid-19 on chest radiographs trained and tested on a large us clinical data set. *Radiology*, 299(1):E167–E176, 2021. 5

[55] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[56] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020. 8