A Method for Evaluating the Capacity of Generative Adversarial Networks to Reproduce High-order Spatial Context

Rucha Deshpande, Mark A. Anastasio, Senior Member, IEEE, and Frank J. Brooks

Abstract—Generative adversarial networks (GANs) are generative models with the potential to revolutionize biomedical imaging. The overarching problem with GANs in clinical applications is a lack of adequate or automatic means of assessing the diagnostic quality of images generated by GANs. We demonstrate several statistical tests of images output by two popular GAN architectures. We designed several stochastic object models (SOMs) of distinct features that can be recovered after generation by a trained GAN. Several of these features are high-order, algorithmic pixel-arrangement rules which are not readily expressed in covariance matrices. We designed and validated statistical classifiers to detect the known arrangement rules, and tested the rates at which the GANs correctly reproduced the rules under a variety of training scenarios. We found that ensembles of generated images can appear accurate visually, and correspond to low Fréchet Inception Distance, while not exhibiting the known spatial arrangements. Additionally, GANs trained on a spectrum of distinct spatial orders did not respect the order prevalence given in the training data. The main conclusion is that while low-order ensemble statistics are largely correct, there are numerous quantifiable errors per image that plausibly can affect subsequent use of the GAN-generated images.

1 INTRODUCTION

GANs in medical imaging, stochastic object models, statistical image analysis, generative model evaluation

1.1 Demonstration of the problem

An illustration of the problem described above is provided through radiomic feature analysis of the NYU fastMRI Initiative [22] brains dataset. Radioimic features are often employed for decision support and treatment planning in clinical scenarios [23], [24]. Any GAN deployed in a medical imaging pipeline would be expected to reproduce the radiomic feature distribution of the training ensemble such that the clinical utility, or the derived clinical outcome based on a GAN-generated ensemble, is unchanged from that of the training data. As shown in Fig. 1, the principal components of extracted radiomic features indicate distinct feature clouds for images from the true clinical images and those generated from GANs trained on them.

In fact, these differences are evident for every feature family given by PyRadiomics [25], a popular software package for extracting standard radiomic features. The differences are so great that a statistical classifier using these features as predictors has nearly 100% accuracy at distinguishing training images from generated images. Thus, the true radiomic feature distributions were not maintained by the GAN. This elucidates the urgent need for a systematic exploration of the capacity of generative models to reproduce clinically relevant constraints on the occurrence of image features.

1.2 Overview of the proposed methodology

Each new realization produced by a trained GAN is random. Thus, it is not possible to know the “intent” of the GAN or
compute an error as a difference between new realization and known truth. This is the same problem encountered in image science during the development of an imaging system. When the expected truth is not available for each output, a stochastic object model (SOM) might be helpful. An SOM is a model with some capacity to describe variance in an object to be imaged, or, in the present case, learned. One very simple example SOM is the so-called lumpy model [29] where distributions are specified for the number, radii, location, and orientation of two-dimensional Gaussian functions. A new realization of the SOM is then made by drawing a parameter vector from the prescribed distribution and superposing Gaussian functions specified by those parameters. If a GAN where trained on an ensemble of lumpy realizations, a post-hoc image processing could be applied and each of known parameters estimated. This scheme would give one measure of the quality of the generated outputs, however, it also could unfairly include a substantial error due to the post-hoc parameter estimation. That is, even “real” realizations of the lumpy SOM could lead to very poor post-hoc parameter estimates. For these reasons, we carefully designed SOMs to be both rich in features relevant to the analysis of diagnostic images and have those features readily recoverable via post-hoc image processing.

The proposed work addresses the problem of evaluating GANs via purposefully designed SOMs that encode specific spatial arrangements at specified prevalences, i.e., “high-order” information. Here, higher-order information is any contextual information not readily expressible in, or detected via, grayscale histograms or covariance matrices. For example, consider a hypothetical chest radiograph of a single patient that exhibits one fewer pair of ribs than it should. That anatomical deficiency might not induce a detectable error in any known summary statistic, however, a human physician can know instantly, from external knowledge, that the image is suspect. In this case, whatever unspecified statistics adequately describe the missing ribs would be considered as higher-order statistics which reflect the additional contextual information. In this way, high-order statistics convey spatial and contextual information and thus should not be confused with high-degree moments of a probability distribution function.

Another way of encoding external information is in the prevalence of features or groups of features. For example, a human heart as seen in a radiograph might be thought of as a set of visual features that should occur precisely once whereas ribs are a different set of features that occur with greater prevalence. Furthermore, these two feature sets are coupled in that a subject with a given size chest ought to have a constrained range of heart size. In other words, the joint probability of whole-feature occurrence itself can be a higher-order statistic.

In this work we demonstrate that external knowledge can be built into training images algorithmically, such that that knowledge can be verified readily after generation, and without explicitly specifying formulas for any particular statistic. This means that we have a ground truth for testing GAN-generated images for various contexts. We then employ several distinct SOMs in several experiments to assess the extent that GANs learn higher-order information along with whatever low-order information is learned during training.

2 Methods

2.1 Radiomic feature analysis

Slices were extracted from volumes in the fastMRI brains dataset (T2 contrast; 3T magnet strength) employing custom code written in Python v3.7. To mitigate the effect of brain size, only slices in which the area of the brain is greater than half of the maximum brain area found within the entire ensemble were included. In order to identify the brain region, masks were generated by the following order of operations: grayscale intensity saturation at the 99th percentile, gaussian blurring with $\sigma=1$, thresholding to the maximum intensity in an ROI of width 40 pixels along the vertical edge of the image, hole-filling, and selection of the largest foreground area in the image. Images were then resized to 256x256, normalized and saved as 8-bit images, yielding a dataset of 17357 slices in all.

PyRadiomics v2.2.0 was employed for radiomic feature extraction from both the original and GAN-generated images. Only the foreground area, as defined by masks previously described, was selected for feature extraction. These image data were quantized into 32 grayscale levels and the following settings specified: co-occurrence distances 1-3 pixels and Gaussian kernel widths $\sigma=1.2$ for all 2D features in the library, giving a total of 1023 features extracted for each image. Principal component analysis was performed separately on each of two datasets, consisting of real and GAN-generated images.

2.2 Description of SOMs

All realizations from all ensembles described below are 8-bit, size 256x256-pixel images. Sample realizations from each ensemble are shown in Fig. 2. Ensembles from the designed SOMs have been made available on Harvard Dataverse: https://doi.org/10.7910/DVN/HHF4AF.
The clustered lumpy background (CLB) [27] is a well-established SOM for the objective assessment of image quality [27], [28]. Realizations consist of randomly generated clusters of Gaussian lumps at various scales and orientations drawn from predetermined distributions. Images generated from this model are expected to be visually similar to mammographic images and exhibit similar statistics up to the second order [27], which, here, is still low spatial order. For the present work, the default parameters [27] of this model were maintained except the mean number of clusters, which was scaled according to the image area as described in [27] to generate 256x256 realizations. The training ensemble comprises 40000 such realizations.

### 2.2.2 Flags SOM

We designed the eight-class flags SOM for testing the joint reproducibility of pre-specified, first-, second-, and high-order image features at once. A 256x256 pixel image of all zeros is first delineated into a grid of 16x16 pixel squares. At prescribed locations, shown in Fig. 3, entire blocks are declared to be foreground. An image class is then one of the eight distinct foreground arrangements.

All classes comprise exactly 80 foreground squares and thus 176 background squares; this eliminates the zero-order variance in the number of pixels of interest. 80x256 variates are chosen from a scaled Beta distribution with parameters $\alpha = 4$, $\beta = 2$, $\lambda = 96$, $\sigma = 152$, and arranged at random within only the foreground squares. Similarly, the remaining 176x256 background pixels are chosen from another scaled Beta distribution with parameters $\alpha = 2$, $\beta = 4$, $\lambda = 8$, $\sigma = 192$ and then also are randomly arranged in the background. The intensity range of the variates is 97-247 for the foreground and 9-199 for the background. The result is an image with foreground brighter than the background where the image intensity follows a mixture distribution:

$$80/256 \text{Beta}(4,2,96,152) + 176/256 \text{Beta}(2,4,8,192).$$

Also of note is that there are 24 certain squares that are never foreground, in any class; this is analogous to anatomical constraints in location for a clinical feature. Together, these classes enable a variety of experiments for exploring how much of each informational order the GAN learns. For example, the extent that learning the correct foreground structures and random arrangements (second-order) also means learning the correct grayscale intensity distributions (first-order) while never misplacing a foreground square in a forbidden location (high-order) can be tested. By choosing the frequency of each distinct foreground structure, the ensemble class prevalence can also be tested—this is equivalent to testing whether the distribution of a clinical feature set is maintained. The flags ensemble consists of 65536 realizations per class.

The blocked nature of the images makes for easy post-hoc classification, performed as follows: to identify the foreground in a GAN-generated image, the realization is split into 16x16 tiles, each tile is identified as background or foreground by comparing its intensity mean against a threshold of 140 chosen from the mixture distribution to be halfway between the modes. Class is then determined based on template matching through the computation of the mean absolute error between the identified foreground blocks and the known class template blocks. Perfect template match is achieved for 96% and 98% realizations from the ProGAN-, and StyleGAN2-generated ensembles. In each respective case, the maximum error in the ensemble corresponds to less than 1/3rd, and 1/4th of the blocks in a realization being wrongly classified. Realizations that did not perfectly match the original class templates were excluded from further analysis and several of those retained were visually spot-checked to ensure that they were well-formed. This abates the effect of foreground formation error and post-hoc processing on the statistical analysis of generated realizations.

### 2.2.3 Voronoi SOM

The Voronoi SOM, also an eight-class SOM, enables testing of second- and high-order information from randomized sets of image features. Class is determined by the fixed number of Voronoi regions within each realization. These range from 12 to 96, in increments of 12. Within each realization, region centers are placed in a spatially random manner, unlike the fixed foreground locations in the flags SOM; this provides an additional source of object variance. Furthermore, all pixels in a region are allocated a single grayscale intensity drawn from a set of 128 predetermined, equidistant values between 0 and 254. Most importantly, the grayscale intensity increases monotonically with area, which is a high-order feature. This SOM is representative of clinical images, such as microscopy or histology images, with multiple, positionally independent regions of interest within an image, each having a distinct intensity. Like the
flags SOM, the Voronoi SOM also allows for testing the ensemble class prevalence, but with feature sets at multiple spatial scales, simultaneously. This multi-class ensemble consists of 65536 realizations per class. For the analysis of GAN-generated images, post-processing involves region detection to determine class followed by the extraction of region-wise grayscale values. To identify region boundaries, each realization is converted to a binary image by thresholding with the largest eigenvalues of the Hessian matrix followed by morphological erosion and skeletonization for further cleaning of boundaries. The regions are then identified through flood-filling and fine-tuned via the watershed transform. It is noted that although this method of region detection is not perfect, it is still sufficiently robust for the experiments proposed.

2.2.4 Alphabet SOM

The alphabet SOM is a single-class ensemble designed to test the per-realization prevalence of features and feature-pairs, independent of spatial scale and grayscale intensity. As discussed in a later section, per-realization feature prevalence is a fundamental property of clinical datasets as it corresponds to testing for anatomical plausibility. Each realization in the alphabet SOM consists of 64 letters of equal size, drawn from a set of 8 letters at predetermined, per-realization frequencies. These are exactly (frequency-letter): 16-H, 2-K, 1-L, 4-V, 8-W, 1-X, 8-Y, 24-Z in each realization. The letters are rastered 8x8 on a black background (refer Fig. 1). This ensures that the effects of grayscale intensity, feature size, and feature location is uniform across realizations, and rules of prevalence can be tested independently.

In addition to the external knowledge represented through per-realization letter prevalence, two rules representing the prevalence of strongly and weakly paired features are also built into each realization. First, the letters in the pairs H-V and W-Y occur exactly 4 and 8 times within a realization respectively, such that the second letter must follow the first. Second, the letters H and Z occur as H-Z or Z-H pairs with equal probability, and at least 12 such pairs occur within each realization. Greater number of H-Z pairings can occur by chance. It is also noted that a letter pair could be split across two rows if the first letter in the pair appears at the end of a row; thus, the second letter would follow at the beginning of the next row. The alphabet ensemble consists of 131072 realizations. For the analysis of GAN-generated images, each realization is split into 32x32 pixel blocks. The letter within the block is identified by template matching against the original letter templates. To quantify the uncertainty in the identity of each letter, a 16-unit scale is established spanning the range of the mean absolute errors computed for all pairs of letter templates. A score of 12 or higher on this scale was found to correspond to visual ambiguity in letter recognition, and was hence chosen as a cutoff for the exclusion of badly formed letters from further analysis.

2.3 Network trainings

Two popular GAN architectures: ProGAN [29] and StyleGAN2 [30] were employed for this work. There are several hyper-parameters available for tuning a GAN. In almost all cases, these were chosen as prescribed by the authors in the original work or code. The only exception was the ProGAN training schedule for the CLB, flags and Voronoi datasets. Training was performed until 7 million images were processed by the discriminator (transition set to 400k images); this training schedule was found to be sufficient in terms of visual quality, FID scores and loss curve convergence. However, convergence occurred much later in two other cases: alphabet and fastMRI brains datasets, and hence, to ensure sufficient training, these were trained until 12 million images were processed by the discriminator, using the default configuration of StyleGAN2(config-e) [30]. Trainings were performed such that the discriminator was shown 4 million images for all cases except the CLB, for which this number was 7 million because of later convergence as determined by the loss curve. The regularization parameter $R_1$ was set to 100 and the truncation parameter $\psi$ was set to 0.5; both are default values for the chosen configuration. The trainings were performed on Nvidia GeForce GTX 1080Ti, 1080 and Tesla V100 GPUs and typically took between 4 and 14 days per training on a single GPU. A total of 10240 realizations, for each dataset, were generated from each network for further analysis.

It is explicitly noted that the goal of this work was not to achieve the best possible performance of any network, but simply to demonstrate the utility of the designed SOMs for assessing common GANs that are trained in a typical way.

3 Results

3.1 Reproducibility of intensity distributions

Ensemble grayscale intensity distributions of all training datasets, namely: fastMRI, CLB, flags, Voronoi and alphabet, were compared against those of the corresponding GAN-generated images and were found to overlap significantly. The greatest differences were observed for the fastMRI brains dataset (Fig. 4, inset). The quantile-quantile plot (Fig. 4) for this dataset demonstrates the deviations in the grayscale quantile values of the GAN-generated images from the corresponding quantiles of the clinical training data. The imperfect replication of the intensity distributions implies that the value of at least some measurements and features must be wrong in some cases.

3.2 Results from the flags SOM

The ensemble grayscale intensity distributions of GAN-generated images from both networks nearly perfectly matched the prescribed mixture distribution of the training dataset. As described in Sec. 2.2.2, the foreground and background in the GAN-generated images were first identified with a post-hoc classifier. The foreground structures were largely well-formed for ensembles from both networks. In the occasional malformations of the foreground (≈1 in 20 in the worst training cases), blending of two templates was observed, however, the forbidden foreground locations were always respected. The observed intensity distributions for the foreground and background were tested separately against the prescribed distributions. A realization was considered acceptable if the $\chi^2$ statistic computed on either the
foreground or background intensity distribution was below a generous tolerance set to the 99.5th percentile of the distribution of the $\chi^2$ statistic for the corresponding intensity distribution of the training data. None of the foreground distributions from the GAN-generated realizations from either network were within tolerance. About 63% and 94% images exhibited background distributions outside tolerance for ProGAN and StyleGAN2 respectively. This indicates that although the grayscale distribution was well replicated in the GAN-generated ensemble, the specific distributions of the foreground and background often were not reproduced per realization. This further indicates that first-order statistics, which are computed from the foreground and background intensity distributions, could fail to match those of the training data. Such a failure not only implies that the distinct foreground and background intensity distributions, such as those for white matter or gray matter in a medical image, are not learned, but also that the application of a statistical observer or post-processing task such as thresholding or segmentation, could be adversely affected.

Next, the reproducibility of the prescribed randomness in pixel arrangement, which is representative of second-order features, was tested via the tile-wise computation of Moran’s Index ($I$) of spatial autocorrelation [31] for each of the 256 tiles in a realization. A realization was considered acceptable if less than 3 tiles in a realization violated the randomness prescribed in the training distribution. Here, 3 tiles corresponds to the 95th percentile of $I$ values, assuming an average error of 1/number of tiles per realization. This acceptance threshold of 3 tiles per realization could be modified according to the relevant tolerance for a task. On average, 4% and 10% of the realizations violated the distribution of $I$ for the foreground and backgrounds for ProGAN, while the proportion is about 6% for both subsets for StyleGAN2. These results imply that a majority of the realizations in ensembles generated from either network reproduce randomness in pixel arrangement. However, a non-negligible proportion, up to 1 in 10, of the realizations do not exhibit the prescribed randomness, and thus, any inference based on the presumption of randomness could be incorrect. Similarly, the performance of an objective numerical observer [32] also could be negatively impacted.

We also observed that the mean class prevalence was found to match the expected mean of 1/8, corresponding to uniform prevalence of the 8 classes in the training ensemble. Although the standard deviation was negligible for ProGAN ($\sigma=0.8\%$), it was non-negligible for StyleGAN2 ($\sigma=10\%$), indicating that some classes were preferentially generated in the latter case. For a clinical dataset, the relevant prevalence might not be reproduced in a GAN-generated ensemble and thus, have significant implications when employed for data augmentation or statistical power calculations.

Thus, for this one SOM, both second-order features and the per-image prevalence of second-order features was reasonably well reproduced; however, the first-order information per-image was essentially always wrong, even though the ensemble mean intensity distribution appears correct.

### 3.3 Results from the Voronoi SOM

The Voronoi SOM represents feature sets at multiple spatial scales while relaxing the positional constraint of the feature of interest as in the flags SOM. First, it was observed that
low-amplitude artifacts were present within each region of each generated image. Each region in the training data is a unique constant value but the Laplacian zero crossings (see Fig. 5 top row) reveal high-frequency structure in the generated realizations. Such artifacts, possibly characteristic of the convolutional network architecture, could affect clinical decision-making because the original second-order information—and thus, possibly any derived texture statistic—is not consistent with the original dataset.

Other artifacts, more visually apparent, constitute curved region boundaries or the presence of a weak boundary in a region of approximately constant intensity. The presence of such artifacts can confound any classifier or analysis that is calibrated on the original dataset (see Fig. 6). Furthermore, within each realization, the high-order rule that demands that the shade monotonically increases with area was not reproduced. The expected Spearman rank-order correlation was lower in the GAN-generated images ($\rho =0.8, 0.7$ for ProGAN and StyleGAN2 respectively) as compared to the true dataset ($\rho =0.9$) as shown in Fig. 7. It is noted that $\rho =0.9$ and not exactly 1, due to some error in the post-hoc delineation of regions. If the grayscale intensity represents a physical property, such as tracer uptake, a breakdown of the correlation between area and intensity implies that the realizations have at least partially lost their quantitative meaning. Last, additional high-order information, encoded as the class-specific number of tiles in a realization, was tested via ensemble class prevalence. It is noted that the number of tiles per realization, or class, ranges from 12 to 96, in increments of 12, in the original ensemble. Although the post-hoc classifier underpredicts the high classes, overlap in the neighboring classes is minimal except for the two highest classes; this error is not large enough to change the conclusion of the class prevalence experiments described below. In the first experiment, all eight classes were represented equally in the training ensemble. As shown in Fig. 7, the prescribed class prevalence within the ensemble was not reproduced, even after accounting for the error of the post-hoc classifier.

Within a generated ensemble, the high-frequency classes were found to have very low prevalence, which also was confirmed visually. Additionally, the more high-frequency images, i.e., classes 84 and 96, were less well learned and exhibited more errors in region formation. This also was verified visually from a 100 random samples from each ensemble to ensure that these errors were beyond that of the classifier. Next, preservation of distinctive classes in the generated ensemble was assessed via the following five experiments. In two experiments, the following subsets of classes were excluded from the training ensemble: (i) (36, 48), (ii) (60, 72) , while in the other three, only the following subsets were included (iii) the two extreme classes- (12, 96) (iv) single class 36 (v) single class 72. Five separate networks were trained for each of these cases per architecture, i.e., ProGAN and StyleGAN2. To determine class after region detection on the generated ensembles, the distribution of the number of regions per realization over the ensemble generated by networks trained with equal class prevalence was binned into eight distinct classes, and treated as reference, separately for each network. In all cases, the GAN-generated ensembles did not maintain the class prevalence of the training data (see Table 1). Examples of realizations belonging to classes excluded from training were observed in the ensemble, indicating interpolation in the learned feature space. In addition, some realizations from the ProGAN-generated exhibit multiple classes at once (see Fig. 5), within different regions of the same realization. For example, within a single realization, the number and size of the regions in the top right quadrant might be more consistent with class 72 while in the bottom left quadrant, it is more consistent with class 36. In a clinical dataset, this would represent mixing of multiple, distinct pathologies to create unrealistic pathologies, which could confound clinical decision-making.

### 3.4 Results from the alphabet SOM

Although most letters appear well-formed, errors occasionally were observed as seen in Fig. 8. A sample realization and its corresponding uncertainty map are shown in Fig. 8A. As described in the Methods, error here is the pixel-wise difference from the known letter templates. Typically, the error rates were: 1 in 6400 letters for ProGAN and 1 in 128 letters for StyleGAN2 respectively, indicating that almost all letters in a realization were visually recognizable. It is also noted that the post-hoc classifier assigns an identity to all letters but only visually recognizable letters are retained.
TABLE 1
Class Prevalence in the GAN-generated Ensembles (P: ProGAN, S: StyleGAN2).

| Classes included in the training ensemble | Class prevalence in the generated ensemble (%) |
|-----------------------------------------|----------------------------------------------|
|                                         | P | S | P | S | P | S | P | S | P | S | P | S | P | S | P | S |
|                                          | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1,2,5,6,7,8                             | 14| 10| 14| 18| 7 | 12| 15| 14| 26| 23| 17| 4 | 6 | 0 | 0 | 0 |
| 1,2,3,4,7,8                             | 14| 19| 22| 34| 21| 22| 12| 13| 14| 8 | 14| 3 | 3 | 1 | 0 | 0 |
| 1,8                                     | 44| 53| 2 | 1 | 2 | 1 | 3 | 3 | 14| 13| 28| 21| 7 | 8 | 0 | 0 |
| 3                                       | 0 | 1 | 24| 22| 76| 73| 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6                                       | 0 | 0 | 0 | 1 | 3 | 7 | 30| 20| 57| 44| 10| 25| 0 | 3 | 0 | 0 |

For classes not included in training, class prevalence should be exactly zero. This is not the case due to interpolation and extrapolation of features across classes for both networks.

Fig. 9. Results from the alphabet SOM. Left: Distribution of the chi-square statistic computed for the per-realization single-letter frequencies shows non-negligible errors in case of both networks. Right: The expected paired-letter prevalence (H-V, W-Y = 4, 8) is not respected by either network. A wide range of values is seen for both networks indicating that perfect prevalence is achieved only by chance.

(see Methods). Thus, in all subsequent analysis, a slightly malformed but clearly recognizable letter is considered no different than a perfectly formed letter. This abates the effect of minor feature shape variance from further analysis. Although the GAN-generated ensembles show high visual similarity with the training ensembles, the high-order rules of feature prevalence were explicitly tested as described below.

The observed frequency of letters was compared to the prescribed frequencies given in Sec. 2.2.4 via the $\chi^2$ goodness-of-fit test. Only 351 ProGAN and 228 StyleGAN2 realizations were found to be outside the 95% critical value of the chi-squared test. However, this means only that the letters appear to have been drawn from the prescribed distribution, not that a realization is correct. In fact, we observed that only 2 ProGAN realizations and 6 StyleGAN2 realizations had the exactly correct letter frequency; recall, this frequency is identical in every training image. Thus, only by rare chance was any realization correct in high order. In a clinical context, if natural variation exists in the prevalence of an anatomical feature, most realizations would be acceptable. However, if the anatomical feature prevalence is exact, e.g., the number of bones in a joint, it is possible that essentially none of the realizations from a GAN-generated ensemble might be acceptable.

Next, the per-image prevalences of unordered and ordered letter-pairs were tested. The unordered pair H-Z, which has an expected frequency of at least 12 per realization, occurred at the following frequencies (mean ± standard deviation) for the training, ProGAN- and StyleGAN2-generated images respectively: 19±2, 19±4 and 19±4. This indicates that although the mean frequencies are the same, the variance found in the generated images extends beyond that of the original dataset and hence, many realizations in the generated ensemble violate the training distribution of the feature pair. The fixed ordered pairs H-V and W-Y were expected to occur at frequencies of precisely 4 and 8 respectively, but were observed to occur at a wide range of frequencies for both networks as seen in Fig. 9.

The single letters V and Y which never occur without the partner in the training data, occurred without the other member of the pair up to 100% of the time. Hence, pairs of clinical image features that may be expected to have anatomically known, relative locations and prevalence might not appear in a GAN-generated ensemble.

Last, to test for positional bias in letter prevalence, the expected per-realization distribution of letters at each position computed from the training ensemble was compared against the corresponding observed frequencies in the GAN-generated ensemble via the $\chi^2$ goodness-of-fit test. As seen in Fig. 10 this statistic shows non-random spatial distribution of errors. Some of this behavior could be attributed to the separation of paired letters at the end of a row that might not be captured by a convolutional window, causing “streaks” of error such as the one observed in case of StyleGAN2 at the right of Fig. 10. This indicates that more errors in feature prevalence might manifest in certain locations of a realization than in others, and that such error trends might be network specific.
Therefore, not only do the generated images not have the correct high-order arrangements but the errors are not distributed randomly across a given realization. Thus, “visually clinical-like” GAN-generated images might have diminished clinical value due to an unrealistic anatomical or pathological representation.

3.5 FID-10k scores

Although all trained models achieve reasonable FID-scores, convergence in loss curves, and high visual similarity of the generated images with the corresponding original dataset, the proposed SOMs demonstrate that a majority of the realizations from these ensembles are still incorrect in terms of per-realization feature prevalences. FID-10k scores for all datasets and both networks are presented in Table 2. The scores are consistent with typical network trainings and visual image quality. In all cases, the last model before end of training was chosen and it is possible that another instance of the network in the same training trajectory might correspond to a lower FID score than that of the chosen model. Alternatively, many other training hyper-parameters could be tuned to possibly yield a lower FID score. Hence, no claims are made about the relative performance of the two architectures or that the chosen models were the best instances from the specific network architecture in terms of FID.

4 DISCUSSION

Although much improvement has occurred in the realism of GAN-generated natural images \[33\], evaluation of these methods for deployment in domains where domain expertise is inextricably tied to image perception, such as in medical imaging, still remains a challenge. For example, an unnaturally shaped organ in a GAN-generated medical image can be immediately spotted by a veteran radiologist but might not be so perceived by other observers, human or numerical. To partially circumvent this challenge, the proposed SOMs provide a method for encoding high-order information relevant to medical imaging while also allowing the recovery of this information from GAN-generated ensemble. The availability of such tests enables the user to make an informed choice about the network architecture for a given task. The capacity of a network architecture can be evaluated in this context, and potential clinical risk arising due to the absence of such knowledge can be assessed.

While we understand that each instance of a particular GAN is unique, and thus that the results of one instance likely do not apply directly to another, any instances trained from the same architecture likely share some common learning capacity. We speculate that if an additional term representing a norm between expected and per-realization statistics and/or feature prevalences could be built into the discriminator, and weighted at an appropriate time during training, a GAN might better represent multiple orders of information at once and thus be of greater clinical utility. We envision the use of designed SOMs as a kind of “necessary but not sufficient” triage of GAN capacity. Our supposition is that if a particular GAN demonstrably fails at recovering the fundamental image properties one prescribes—such as grayscale intensity distribution, spatial randomness, and pre-specified feature prevalences—then that architecture could fail to accurately reproduce any sort of diagnostic image that comprises those fundamental properties. This is why non-medical object or image models such as the ones we have designed can be relevant to estimating the probability that a GAN has made errors in medical images.

4.1 Relevance to clinical scenarios

In medical images, features can have quantitative, structural, and positional significance within each realization; this can be partially described by statistics spanning multiple orders of information. However, as observed in the results from the flags SOM, the joint reproduction of statistics across multiple orders might be a challenging task for the chosen GANs. Specifically for the flags SOM, the prescribed intensity distributions for the foreground and the background were not reproduced for over 60% of the ensemble, although spatial randomness and foreground structure were not reproduced in only about 10% of the ensemble. Thus, while some orders of information are correctly replicated, other orders might not be and, hence, the ultimate utility of a generated realization might be determined by the order of information required for a specific diagnostic task. For example, if a GAN were to be employed for simulating a PET-imaged dataset of a certain tumor type, a majority of the GAN-generated tumors could be correct in shape and structure but significantly different in the expected intensity distribution and texture. Thus, many radiomic features computed on the generated ensemble could be wrong. Drawing clinical or diagnostic inferences from a generated ensemble, even when employed alongside the original data for its augmentation, might translate these sorts of GAN-induced errors to false clinical predictions and prove potentially dangerous for patients.

To further explore the reproducibility of high-order information, the Voronoi SOM was designed such that image features, corresponding to the Voronoi regions, were ergodic. An analogous clinical example is a histology image, depicting multiple cell types, each with characteristic textural features and staining intensity but able to appear anywhere in the field of view. These physical characteristics must be correct if generated realizations are to be considered diagnostically accurate. As described in Sec.3.3, interpolation of features across the spectrum of spatial scales was observed within single realizations as well as across the

| Table 2 |
|-----------|
| FID-10k Scores for Both Networks on All Datasets. |
| Dataset | ProGAN | StyleGAN2 |
|----------|---------|-------------|
| fastMRI  | 8.1     | 24          |
| CLB      | 24      | 22          |
| Flags    | 13      | 29          |
| Voronoi  | 16      | 27          |
| Alphabet | 5.5     | 8.9         |

Low FID scores are not indicative of reproducibility of high-order per-image statistics, which have to be assessed independently.
ensemble. For augmentation purposes, this might be good if the learned feature distribution does not violate that of the population itself, but for classification, a violation of the expected feature distribution might prove disastrous. Another potential factor causing a negative impact on a downstream classification task is the presence of low-amplitude artifacts, such as those seen in Fig. [5]. This is an example where the generated images appear largely correct visually but are completely different to a machine learner that is sensitive to derivatives. In addition to these errors in spatial structures, the high-order information encoded as the rank-correlation between area and grayscale intensity also was not reproduced exactly. Thus, the quantitative information, generally representative of physical tissue properties observed via a certain imaging modality, can be unreliable, and the textural features suspect, for the GAN-generated realizations we observed.

Besides feature interpolation or class-mixing observed within single realizations, the class prevalence in the training ensemble is also not correctly reproduced for both multi-class SOMs: flags and Voronoi. Thus, if these particular instances of GANs were used to replicate a clinical dataset for virtual clinical trials, or to generate a training ensemble for a segmentation or classification machine under development, the prevalence of the input pathologies would not be maintained. Most significantly, this bias could be characteristic of the network-architecture and thus, would have to be quantified for each architecture separately.

To assess only the reproducibility of high-order features, and eliminate the effects of position, structure and shading, the alphabet SOM was designed with known per-realization prevalence of single and paired features. Because anatomical plausibility can be represented (at least partially) as the joint, per-realization prevalence of naturally occurring features, it is paramount that this prevalence is maintained within each realization and not just on average, over the ensemble. If a generative model is designed to maximize similarity over the ensemble alone, per-realization errors might be widespread. This was observed in the GAN-generated images of the alphabet SOM, where less than 0.1% of the ensemble had perfect feature and feature-pair prevalence, even when high similarity was indicated not only by ensemble metrics such as the FID score and the $\chi^2$ statistic, but also from visual inspection. Such visually realistic GAN-generated images with incorrect per-realization feature prevalence might have reduced diagnostic value. This could translate to clinical decisions as bias or even a complete failure to learn.

4.2 Applications of SOM-based GAN-evaluation

There is considerable interest in employing GANs for a variety of medical imaging and image analysis pipelines [2], [3]. Augmentation of various medical image data is one key area of interest [34], [35], [36] because the number of unique subjects and/or expert annotations often is too small for meaningful statistical analysis or robust machine training [2], [37], [38], [39]. Even if it is presumed that a trained GAN outputs images deemed “good,” an additional concern is whether that GAN maintained the correct numbers of each kind of data. For example, it is possible for a feature set corresponding to one infrequent pathology to be “easier” for the GAN to generate than another set corresponding to a more frequent pathology. In other words, the representation of features within the GAN and clinical prevalence could be correlated in unknown ways. Multi-class SOMs can be created as described previously and such that the difference between classes can be subtle or profound and then the extent to which a given GAN architecture mixes the classes explored via adjusting the frequency of classing in the training data as was described previously.

Another avenue of interest for GANs is medical domain transfer, where image data acquired from one imaging modality might be converted to another [40], [41]. The motivation is that if a GAN could learn to reliably create, for example, the “style” or appearance of positron emission tomography (PET) from conventional computed tomography (CT), PET imaging would immediately become more widely available. In any such transformation, there might be only a partial overlap in relevant information. For example, a structure responsive to x-rays and visible in a CT image might not uptake the radiotracer used for PET and thus, would not be observed in a PET image. In this way, creation of functional features not seen in CT is tantamount to learning to hallucinate [42]. A first step in assessing the hallucination capacity of a GAN might be in adapting the Voronoi SOM to have multiple intensity channels for a single class. Each channel could correspond uniquely to one of the two domains and differ by grayscale shading rules. For example, the shading rule for one domain demands a strong positive rank correlation of the grayscale intensity with the area, while that in the second domain demands a weak or non-linear rank correlation and includes some regions forbidden to be above a certain intensity. The style transfer GAN could then be trained to convert one channel to the other. A separate region detector for each domain then could be designed and applied on the generated images. The extent to which the known intensity correlations and shading rules are respected in the generated images is one measure of how well the particular GAN architecture can transfer one domain to another without also creating unwanted artifacts.

Other interesting applications that require independent validation of GAN-based results are in super-resolution and image denoising. Specific examples are where lower dose CT image acquisitions are to be enhanced to higher-dose image quality [43] and where speckle noise is removed [44]. Maintaining anatomical accuracy, while correctly reproducing the expected textural information of super-resolved or denoised images, is critical in order for the processed data to retain its clinical value. In essence, both of the proposed applications are, in some key ways, about transformation of intensity distributions in a spatially aware way. Models similar to the flags SOM could be helpful here because a successful GAN trained on these models must simultaneously learn a first-order noise model and second-order texture model. To apply the flags SOM in this scenario, the statistically known distribution of noise for a given imaging modality or a dataset, as well as its spatial arrangement (e.g., random), would be specified. The GAN to be tested would be trained to generate denoised outputs, as desired. Analysis of the generate realizations could then be done
separately for the foregrounds and backgrounds, similarly to that described in Sec. 3.2.

5 Conclusion

The main conclusion is that some popular generative adversarial network (GAN) architectures can make impactful, per-realization errors at a high rate even when summary and ensemble measures of training appear reasonable. The main reason that these errors are difficult to evaluate in diagnostic scenarios, in general, is that there usually is not a mathematically specified ground truth or expert labeling for each GAN-generated realization. It is demonstrated how stochastic object models can be purposefully designed to include known high-order contextual information, analogous to anatomical and physiological information, that also can be quantified post-generation and thus serve as a ground truth. This design can be done algorithmically, without actually specifying a formula for any particular high-order statistic. Several such SOMs were proposed and employed in the evaluation of two popular GAN architectures. Across various training and model scenarios, it was found that the tested GANs failed to simultaneously reproduce all prescribed object features, at once, despite low FID-10k scores and obvious visual similarity between training and prescribed object features, at once, despite low FID-10k the tested GANs failed to simultaneously reproduce all features, at once, despite low FID-10k scores and obvious visual similarity between training and generated data. Specifically, numerous per-realization errors occur in: grayscale intensity distribution, spatial arrangement, and, perhaps most impactful, in the frequency of pre-specified rates of feature occurrence. The corollaries are that GANs should be carefully evaluated prior to diagnostic or clinical use and that designed SOMs can serve as a kind of triage before task-based measures of generated image quality are employed.

Acknowledgements

This work was supported in part by NIH awards EB020604, EB023045, NS102213, EB028652, and the Washington University Imaging Sciences Pathway Trainee Fellowship (ST32EB01485505). This work utilized resources supported by the National Science Foundation’s Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign [45].

References

[1] P.-H. Lu, P.-C. Wang, and C.-M. Yu, “Empirical evaluation on synthetic data generation with generative adversarial network,” Proc. Conf. Web Intelligence, Mining and Semantics, 2019, pp. 1–6.
[2] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” Med. Image Anal., vol. 58, p. 101152, 2019.
[3] V. Sorin, Y. Barash, E. Konen, and E. Klang, “Creating artificial images for radiology applications using generative adversarial networks (GANs)—a systematic review,” Acad. Radiol., vol. 27, no. 8, pp. 1175–1185, 2020.
[4] Y. Zhang et al., “Improving CBCT quality to CT level using deep learning with generative adversarial network,” Med. Phys., vol. 48, no. 6, pp. 2816–2826, 2021.
[5] N. K. Singh and K. Raza, “Medical image generation using generative adversarial networks: A review,” Stud. Comput. Intell., vol. 932, pp. 77–96, 2021.
[6] J. P. Cohen, M. Luck, and S. Honari, “Distribution matching losses can hallucinate features in medical image translation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, 2018, pp. 529–536.
[7] A. Higaki, T. Miyoshi, and O. Yamaguchi, “Concerns in the use of adversarial learning for image synthesis in cardiovascular intervention,” Eur. Heart J. -Digital Health, 2021.
[8] I. Goodfellow et al., “Generative adversarial nets,” Adv. Neural Inf. Process. Syst., vol. 27, 2014.
[9] S. M. Astley, W. Chen, K. J. Myers, and R. M. Nishikawa, “Special section guest editorial: Evaluation methodologies for clinical AI,” J. Med. Imaging, vol. 7, no. 1, 2020.
[10] I. Gulrajani, C. Raffel, and L. Metz, “Towards GAN benchmarks which require generalization,” in Proc. Int. Conf. Learn. Representations, 2019.
[11] P. Salehi, A. Chalechale, and M. Taghizadeh, “Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments,” arXiv:2005.13178 [cs.CV], 2020. [Online]. Available: https://arxiv.org/abs/2005.13178
[12] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, “Assessing generative models via precision and recall,” Adv. Neural Inf. Process. Syst., p. 5228–5237, 2018.
[13] A. Borji, “Pros and cons of GAN evaluation measures,” Comput. Vis. Image Underst., vol. 179, pp. 41–65, 2019.
[14] M. F. Naem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in Proc. Int. Conf. Mach. Learn. PMLR, 2020, pp. 7176–7185.
[15] A. Borji, “Pros and cons of GAN evaluation measures: New developments,” arXiv:2013.05996 [cs.LG], 2021. [Online]. Available: http://arxiv.org/abs/2013.05996
[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two-time-scale update rule converge to a local Nash equilibrium,” Adv. Neural Inf. Process. Syst., vol. 30, 2017.
[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” Adv. Neural Inf. Process. Syst., vol. 29, pp. 2224–2232, 2016.
[18] D. Geman, S. Geman, N. Halonquist, and L. Younes, “Visual Turing test for computer vision systems,” Proc. Natl. Acad. Sci. U.S.A., vol. 112, no. 12, pp. 3618–3623, 2015.
[19] M. J. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, “How to fool radiologists with generative adversarial networks? a visual Turing test for lung cancer diagnosis,” in Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, 2018, pp. 240–244.
[20] A. Semiletov et al., “Comparative evaluation of lung cancer CT image synthesis with generative adversarial networks,” in Int. Conf. on Comp. Sci. Springer, 2021, pp. 593–608.
[21] H. Y. Park et al., “Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: Visual Turing test,” JMRI Med. Inform., vol. 9, no. 3, p. e23328, 2021.
[22] J. Zbontar et al., “FastMRI: An open dataset and benchmarks for accelerated MRI,” arXiv:1811.08839, 2018. [Online]. Available: http://arxiv.org/abs/1811.08839
[23] F. Valdora, N. Houssami, F. Rossi, M. Calabrese, and A. S. Tagliafico, “Rapid review: radiomics and breast cancer,” Breast Cancer Res. Treat., vol. 169, no. 2, pp. 217–229, 2018.
[24] Z. Liu et al., “The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges,” Theranostics, vol. 9, no. 5, p. 1303, 2019.
[25] J. J. Van Griethuysen et al., “Computational radiomics system to decode the radiographic phenotype,” Cancer Res., vol. 77, no. 21, pp. e104–e107, 2017.
[26] J. Rolland and H. H. Barrett, “Effect of random background inhomogeneity on observer detection performance,” J. Opt. Soc. Am. A. Opt. Image Sci. Vis., vol. 9, no. 5, pp. 649–658, 1992.
[27] O. Bouwman et al., “Can the non-pre-whitening model observer, including aspects of the human visual system, predict human observer performance in mammography?” Phys. Med., vol. 32, no. 12, pp. 1599–1609, 2016.
[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Representations, 2018.
[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8110–8119.
[30] P. A. Moran, “Notes on continuous stochastic phenomena,” Biometrika, vol. 37, no. 1/2, pp. 17–23, 1950.
[32] H. H. Barrett and K. J. Myers, *Foundations of image science*. John Wiley & Sons, 2013.

[33] P. Shamsolmoali et al., “Image synthesis with adversarial networks: A comprehensive survey and case studies,” *Inf. Fusion*, 2021.

[34] H.-C. Shin et al., “Medical image synthesis for data augmentation and anonymization using generative adversarial networks,” in *Proc. Int. Workshop Simulation Synthesis Med. Imag.* Springer, 2018, pp. 1–11.

[35] Y.-B. Tang, S. Oh, Y.-X. Tang, J. Xiao, and R. M. Summers, “CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation,” in *Proc. SPIE*, vol. 10950. International Society for Optics and Photonics, 2019, p. 109503V.

[36] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy, “Data augmentation for skin lesion using self-attention based progressive generative adversarial network,” *Expert Syst. Appl.*, vol. 165, p. 113922, 2021.

[37] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9242–9251.

[38] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[39] S. Pandey, P. R. Singh, and J. Tian, “An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation,” *Biomed. Signal Process Control*, vol. 57, p. 101782, 2020.

[40] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, and D. Shen, “Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer’s disease diagnosis,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2018, pp. 455–463.

[41] K. N. B. Boni et al., “MR to CT synthesis with multicenter data in the pelvic area using a conditional generative adversarial network,” *Phys. Med. Biol.*, vol. 65, no. 7, p. 075002, 2020.

[42] S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio, “On hallucinations in tomographic image reconstruction,” *IEEE Trans. Med. Imaging*, vol. 40, no. 11, pp. 3249–3260, 2021.

[43] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, “Generative adversarial networks for noise reduction in low-dose CT,” *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[44] Y. Huang et al., “Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network,” *Opt. Express*, vol. 27, no. 9, pp. 12 289–12 307, 2019.

[45] V. Kindratenko et al., “HAL: Computer system for scalable deep learning,” in *Proc. Practice Experience Adv. Res. Comput.*, 2020, pp. 41–48.