PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Evaluating the capacity of deep generative models to reproduce measurable high-order spatial arrangements in diagnostic images

Rucha Deshpande, Mark Anastasio, Frank Brooks

Rucha Deshpande, Mark A. Anastasio, Frank J. Brooks, "Evaluating the capacity of deep generative models to reproduce measurable high-order spatial arrangements in diagnostic images," Proc. SPIE 12032, Medical Imaging 2022: Image Processing, 120321X (4 April 2022); doi: 10.1117/12.2611807

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Evaluating the Capacity of Deep Generative Models to Reproduce Measurable High-order Spatial Arrangements in Diagnostic Images

Rucha Deshpande[a], Mark A. Anastasio[b], and Frank J. Brooks[b]

[a]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, 63130 USA
[b]Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL, 61801 USA

ABSTRACT

Given the recent interest in the role of deep generative models (DGM) in medical imaging pipelines, it is imperative to evaluate the capacity of such models to generate medically accurate images. Popular methods of evaluation of natural images generated using generative adversarial networks (GANs), a type of DGM, are often applied to medical data. Such methods are insufficient to evaluate anatomical realism, representations of which include high-order spatial information. To our knowledge, no test exists for the faithful replication of spatial statistics beyond the second-order. In this work, purposefully designed stochastic object models (SOMs) are proposed to encode predetermined rules governing the prevalence of features within single images, thus encoding known high-order spatial information within each realization. These SOMs are independent of the network architecture being tested and can also be applied to any new architecture that may be proposed. Two popular GANs are trained on these SOM datasets and the generated images are tested for the encoded statistics. It is observed that although ensemble statistics might be well replicated, this is not necessarily true for realization i.e., per-image statistics. Thus, GAN-generated images might not be ready for clinical use. With the proposed SOMs, the rate of image errors and the rate of feature malformation can be quantified for any architecture, while providing one measure of GAN utility in a diagnostic scenario.

Keywords: high-order statistics, deep generative models, stochastic object models, generative adversarial network (GAN) evaluation

1. INTRODUCTION

The use of deep generative models (DGMs), such as generative adversarial networks (GANs), in medical imaging pipelines has been an active area of research recently. Realizations from a DGM represent variates drawn from an unknown high-dimensional distribution that describes the ensemble of training images. Evaluation of these realizations often involves comparison of statistics derived from either the grayscale intensity distribution or pairs of intensities at a fixed distance. In the context of medical images, expert knowledge is required to perceive contextual differences, such as whether an organ is correct in shape, size and general appearance relative to other organs and given a known pathology. Thus, "high-order" spatial statistics are defined here as those conveying the contextual information not readily or adequately expressible via pairwise pixel correlations alone.[1] In other words, here, "high-order" is not to be confused with the high-degree moments of a first-order image statistic such as the skewness or kurtosis. To our knowledge, no objective method exists that provides a direct assessment of the reproducibility of statistics representative of the high-order spatial information in diagnostic images.

The aim of this work is to provide a method for evaluating the ability of a DGM to reproduce specified spatial arrangements within an image. In particular, how global measures of training relate to individual image errors is explored using algorithmically specified rules in the designed SOMs. This is demonstrated on two popular architectures but is not restricted to them. It is further noted that no claim is made about either architecture being superior in any regard because the goal of this work is not to do a comprehensive assessment of all instances of a particular architecture but only to demonstrate the methodology for use of the proposed SOMs.

Send correspondence to Frank J. Brooks. E-mail:fjb@illinois.edu, Telephone: 1 217 333 1867

Medical Imaging 2022: Image Processing, edited by Olivier Colliot, Ivana Išgum, Proc. of SPIE Vol. 12032, 120321X · © 2022 SPIE
1605-7422 · doi: 10.1117/12.2611807

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

2. METHODS

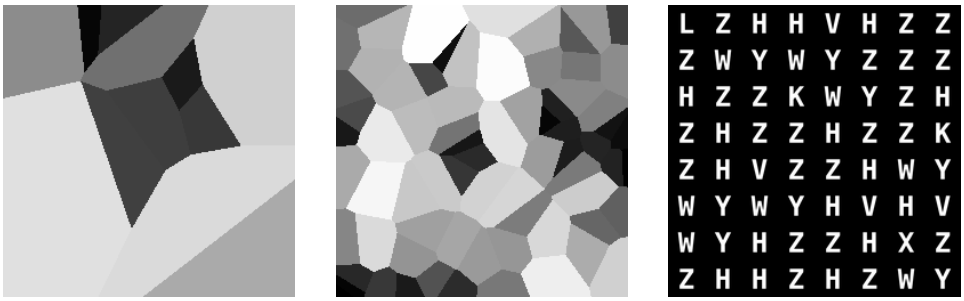
2.1 Reproducibility of radiomic features from a clinical dataset

Two GAN architectures (described in Sec.2.4) were trained on the fastMRI brains dataset.[2] A total of 17357 slices were extracted from volumes with T2 contrast at 3T magnet strength. Slices were selected such that the area occupied by the foreground was at least half of the maximum foreground area over all slices in the ensemble. These were resized to 256x256 and converted to 8-bit after data cleaning. For radiomic feature analysis, PyRadiomics 2.2.0[3] was employed with the following settings: histogram binned to 32 gray levels, distances 1-3 and sigma 1-2, for all 2D features and image classes available in the library, giving a total of 1023 features.

2.2 Designed SOMs

The first SOM, henceforth referred to as "Voronoi", is a set of eight classes of varying spatial order. Each realization within a class is a Voronoi diagram with a predetermined number of regions ranging from 12 to 96, in multiples of 12. Most importantly, the grayscale intensity of each tile within a realization is rank-correlated with its area such that larger areas have higher grayscale intensity. This represents a high-order arrangement rule that tests learning of information beyond typical second-order correlations. In the second SOM, henceforth referred to as "alphabet", each realization is a panel of 64 equally sized letters placed equidistantly on a zero intensity background. Within each realization, the following 8 letters: Z, H, Y, W, V, K, L, X have fixed prevalence. Here, an additional high-order arrangement rule is that each training realization has exactly 4 H-V and 8 W-Y pairs such that the second letter in the pair always follows the first. This enables the measurement of per-image frequencies as well as ensemble frequencies of the letters.

The Voronoi ensemble consists of 65536 realizations per each of the 8 total classes while the alphabet ensemble consists of 131072 realizations. Each realization is an 8-bit image of size 256x256. Sample realizations for each ensemble are shown in Fig. 1. (All data available at Harvard Dataverse: https://doi.org/10.7910/DVN/HHF4AF)



(a) Voronoi: class 12 (b) Voronoi: class 72 (c) Alphabet

Figure 1: Sample realizations from the Voronoi SOM (a) and (b) show classes with varying spatial order corresponding to the number of regions in the realization; (c) the alphabet SOM has pre-specified single letter and letter-pair frequencies that occur in every realization

2.3 Post-processing of generated images

For the analysis of generated images from Voronoi, a statistical classifier is designed to extract the number of regions and corresponding shades. In the case of alphabets, template matching is employed to recognize letters and provide the corresponding uncertainty in recognition of each letter in the generated images. Image class prevalence and per-image feature prevalence are then computed based on these results, while also accounting for the uncertainty in the post-processing methods.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

2.4 Network trainings

Two popular network architectures: ProGAN[4] and StyleGAN2(config-e)[5] were chosen to demonstrate the use of the proposed tests. For ProGAN, the network was trained for 7M images (with transitions set to 400k images) for Voronoi and 12M for alphabet and the clinical dataset, using the default training scheme. For StyleGAN2 (config-e), all trainings were performed for 4M images. The regularization parameter R_1 was set to the default value of 100 for the chosen training configuration and truncation was set to ψ=0.5. For both architectures, the last model was chosen post-training and 10240 realizations were generated for analysis. The trainings were performed on Tesla V100, GeForce GTX 1080 and 1080Ti GPUs and took between 4 and 14 days per GPU. It is noted that for the purpose of this work, the goal was not to achieve the best performance for a network in terms of a chosen metric but to demonstrate the use of the proposed SOMs as a tool for testing a chosen instance of any architecture.

3. RESULTS

3.1 Reproducibility of radiomic features of a clinical dataset

Distributions of radiomic features for true and GAN-generated images were compared by projecting them on the first two principal components. Distinct point clouds for true and generated data are clearly seen for both architectures (refer Fig.2 for principal component analysis of features derived from local binary patterns of the images. It is noted that although the central tendencies of some radiomic feature families across the ensemble of generated images appear reasonably well-aligned with those of the training data, some individual feature families (e.g., local binary patterns) varied wildly and thus individual GAN realizations can be readily classified as fake: an almost perfect AUC ($\geq$0.99) was achieved by a random forest classifier trained on each feature family separately. The FID-10k scores for ProGAN and StyleGAN2 are 8.13 and 24.03 respectively.

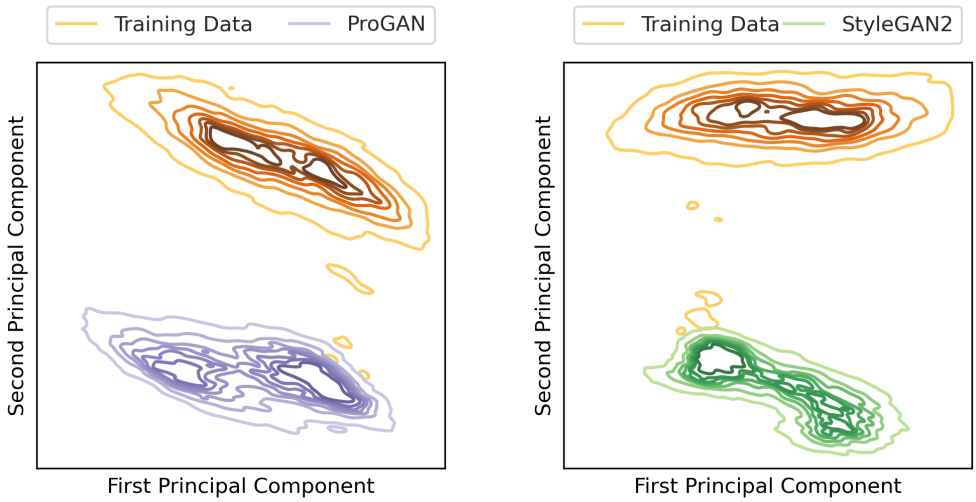


Figure 2: Principal component analysis of radiomic features derived from local binary patterns (LBP) of images shows distinct distributions of real and generated data for ProGAN and StyleGAN2 capturing 86% of the variance each network.

3.2 Results from the Voronoi SOM

As shown in Fig. 3, the equal prevalence of all eight classes in the training data is not respected by either network, even after accounting for the errors in the post-hoc classifier. Empirical testing shows that the specified rank correlation between area and shade in the true data (ρ=0.9) is not maintained in the network-generated ensembles (ProGAN: ρ=0.8 and StyleGAN2: ρ=0.7). In particular, a non-negligible proportion of realizations have poor rank correlation. Lastly, unrealistic realizations with ambiguous class membership are also present in the generated ensemble. Other visually observed errors include presence of high-frequency, low-magnitude

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

artifacts in regions expected to have a constant value (refer Fig. 3), unrealistic curvature of boundaries, and smudged regions. The FID-10k scores for ProGAN and StyleGAN2 are 16.5 and 26.7 respectively.

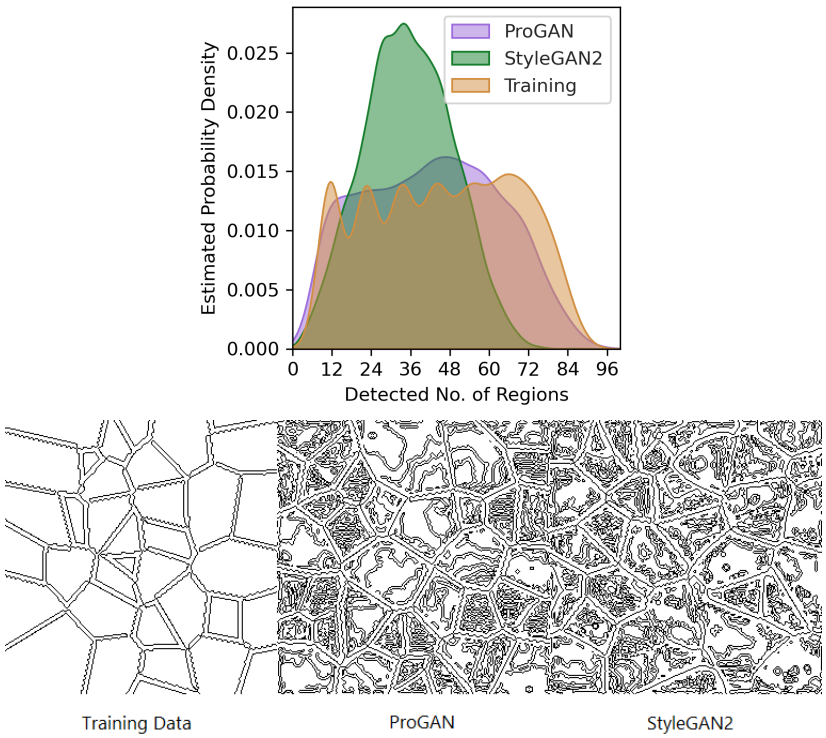


Figure 3: Results from the Voronoi SOM. (Top) Class prevalence in the training ensemble is altered differently by each architecture. (Bottom) High-frequency artifacts are present in the generated images in regions expected to have constant grayscale intensity, visible after computing the Laplacian zero-crossings on realizations from the training, ProGAN-generated and Stylegan2-generated image ensembles (left, center and right panels respectively).

3.3 Results from the alphabet SOM

Statistics upto and including second-order are preserved in the generated images to the extent that individual letters generally appear well formed and are clearly distinguishable, which is also represented in the low FID-10k scores for both architectures (Progan: 5.46; Stylegan2: 8.94). The extent of malformation or uncertainty of a letter can be quantified and letters that are malformed beyond visual recognition are excluded from further analyses (about 1 in 6400 letters for ProGAN and 1 in 128 letters for StyleGAN2). As the expected frequency of each letter in a realization is known, the χ^2 goodness-of-fit statistic is plotted as shown in Fig. 4. The number of realizations beyond the 95% critical value threshold is 203 and 118 for ProGAN and StyleGAN2 respectively within an ensemble of 6000 images, indicating that most images in an ensemble lie within the expected *ensemble* variation. However, the number of "perfect" realizations (χ^2=0) is only 1 for ProGAN and 3 for Stylegan2 within the entire ensemble, suggesting that if compliance with the rule is critical for use of the image, then essentially none of the generated images are acceptable, although the ensemble prevalence of letters is largely respected by both networks. Lastly, the high-order arrangement rule for the occurrence of letter-pairs is tested. The expected frequency for the pairs H-V and W-Y is exactly 4 and 8. However, as seen in Fig. 4, a wide range of frequencies is observed for realizations from both networks. Furthermore, it was observed that the letters V and Y usually occurred without their preceding partner, even this never occurs in the training data. Together, these observations indicate that high-order information rules are not learnt in training.

It is reiterated here that these tests serve as tools for evaluating the capacity of any generative model and no

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

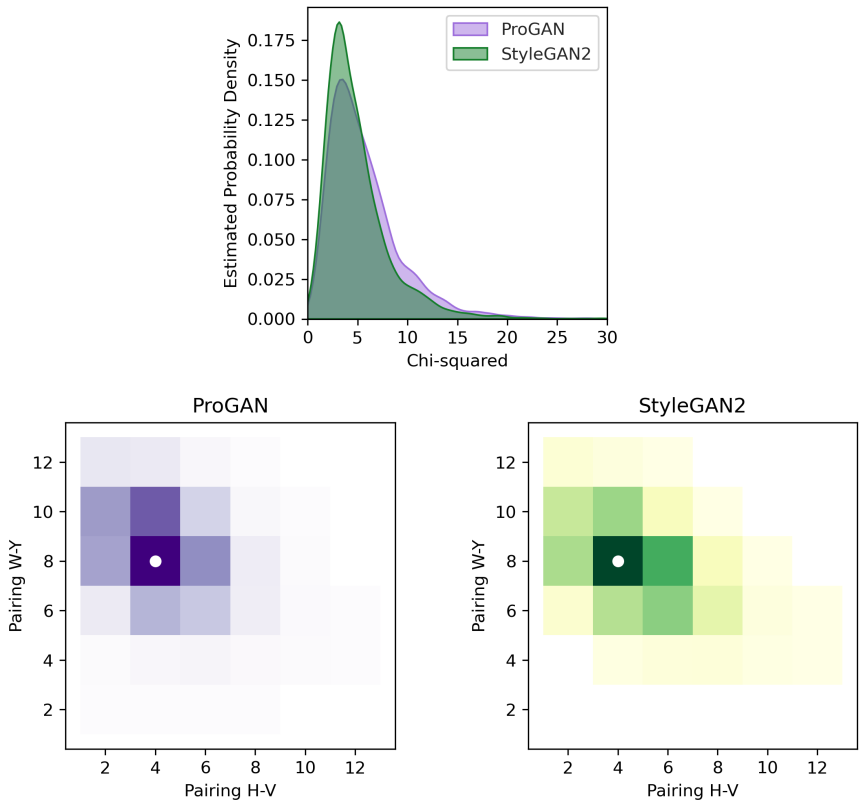


Figure 4: Results from the alphabet ensemble. (Top) Distribution of the chi-squared statistic for single letters per realization in the generated ensembles indicates non-negligible errors in single letter prevalence. (Bottom) Paired letter prevalence indicates that neither architecture reproduces the high-order statistics correctly. Expected prevalence for pairs (H-V, W-Y) is (4,8).

claim is made of the relative performance of the two specific architectures chosen as examples. Only the chosen instances of these architectures are compared. Any new or existing architecture, or even a differently trained configuration of the chosen architectures, could be employed instead.

4. CONCLUSIONS

The altered radiomic feature distributions highlight the need for careful evaluation of the statistical properties of an ensemble generated from a DGM, before it is employed for any medical imaging application. In this direction, the Voronoi SOM serves as an important tool to quantify the capacity of a given DGM to reproduce statistics representative of properties such as the prevalence of distinct pathologies in the training ensemble and their characteristic features. The alphabet SOM allows for assessing individual realizations for properties such as relative shapes and locations of objects and structures (e.g., organs or blood vessels) through per-image high-order statistics, especially when ensemble statistics are respected.

The proposed SOMs provide a general method for the evaluation of reproducibility of high-order information through specific pixel arrangement rules relevant to ensembles of medical images. Such an architecture-independent evaluation of the capacity of a generative model provides more information than conventional tests or FID scores. In the context of the properties tested by the proposed SOMs, an informed choice of a network can be made based on the relative importance of such properties for a given task through comprehensive testing of multiple architectures and their variations. Thus, designed SOMs can be employed for quantifying the utility of a DGM architecture when high-order information is of relevance.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

DISCLOSURE

The work presented here differs from a submitted manuscript in that the latter includes additional stochastic object models and results from an extensive analysis performed after employing these models.

ACKNOWLEDGMENTS

This work was supported in part by NIH awards EB020604, EB023045, NS102213, EB028652 and the Imaging Sciences Pathway (5T32EB01485505). This work utilized resources supported by the National Science Foundations Major Research Instrumentation program, grant 1725729, as well as the University of Illinois at Urbana-Champaign.[6]

REFERENCES

[1] Diaconis, P. and Freedman, D., "On the statistics of vision: the julesz conjecture," *Journal of Mathematical Psychology* **24**(2), 112–138 (1981).

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

[3] Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J., "Computational radiomics system to decode the radiographic phenotype," *Cancer research* **77**(21), e104–e107 (2017).

[4] Karras, T., Aila, T., Laine, S., and Lehtinen, J., "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196* (2017).

[5] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T., "Analyzing and improving the image quality of stylegan," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 8110–8119 (2020).

[6] Kindratenko, V., Mu, D., Zhan, Y., Maloney, J., Hashemi, S. H., Rabe, B., Xu, K., Campbell, R., Peng, J., and Gropp, W., "Hal: Computer system for scalable deep learning," in [*Practice and Experience in Advanced Research Computing*], 41–48 (2020).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 10 Feb 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use